

Alignment for Vision-Language Foundation Models

Zhiqiu Lin

CMU-RI-TR-23-83

Dec 08, 2023



The Robotics Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee:

Deva Ramanan, *chair*

Deepak Pathak

Graham Neubig, *CMU LTI*

*Submitted in partial fulfillment of the requirements
for the degree of Master in Robotics.*

Copyright © 2023 Zhiqiu Lin. All rights reserved.

I would like to thank all my collaborators at CMU and Meta for supporting my research.

To my awesome advisors who supported my research.

Abstract

Recent advancements in vision-language foundation models, exemplified by GPT4-Vision and DALL-E 3, have significantly transformed both research and practical applications, ranging from professional assistance to content creation. These models excel with minimal downstream data and limited human input, primarily leveraging prompt-based interactions. However, aligning them precisely with specific user goals presents a notable challenge. This thesis introduces innovative strategies for improving this alignment. It begins with a novel cross-modal adaptation framework, utilizing textual data to tailor foundational models such as CLIP more effectively to tasks such as visual recognition. It then explores an approach based on ChatGPT for aligning popular proprietary models, like DALL-E 3, to better meet user needs. Lastly, the thesis addresses the challenges in visio-linguistic reasoning, discussing efforts to assess and enhance model fidelity in complex tasks requiring advanced compositional reasoning.

Acknowledgments

I extend my deepest gratitude to my advisor, Prof. Deva Ramanan, for his unwavering support, invaluable wisdom, and insightful guidance. Prof. Ramanan has not only been a mentor but also a pivotal figure in shaping my approach to research. He instilled in me the importance of first-principle thinking and a pursuit of science that seeks understanding beyond mere knowledge. His clarity in dissecting complex problems and creativity in addressing them have profoundly influenced my own research methodology. His commitment to excellence, combined with a relentless pursuit of science, has been a constant source of inspiration and learning for me. I am immensely thankful for his patience, encouragement, and the countless hours he dedicated to my development, both academically and personally.

I am also profoundly grateful to my co-advisors, Prof. Deepak Pathak and Prof. Graham Neubig, for their invaluable contributions to my research journey. Prof. Pathak's exceptional creativity and innovative mindset have been fundamental to the success of numerous publications of mine. His unique ability to approach challenges with out-of-the-box thinking and apply groundbreaking solutions to intricate problems has not only significantly elevated the caliber of my research but has also considerably amplified its impact in the field. His visionary perspective consistently provided fresh angles and insights, greatly enriching the depth and breadth of my work. Prof. Neubig, with his profound expertise in language processing, has played an instrumental role in organizing and spearheading a large-scale, multi-lab project. His depth of knowledge have greatly broadened my understanding and approach to interdisciplinary research. The support and guidance from both Prof. Pathak and Prof. Neubig have been cornerstones of my development as a researcher.

A special word of thanks goes to our industry collaborator, Pengchuan Zhang, for his mentorship and the practical insights he brought to our collaboration. His experience in the industry has been a valuable asset, providing a unique perspective that enriched my research experience. His enthusiasm and practical approach to problem-solving have greatly contributed to my understanding of the real-world applications of our research.

As I continue my journey through the PhD program, the guidance and support of these mentors are the guiding lights illuminating my path. Their

collective wisdom, combined with a shared passion for groundbreaking research, have not only shaped my current thesis but are also instrumental in my ongoing endeavor to advance the state of the art. Their belief in my abilities, their commitment to nurturing my academic growth, and their invaluable insights into the complexities of academia and research continue to inspire and drive my pursuit of excellence in this field.

I am also deeply thankful to my labmates at CMU, who have been an integral part of my academic journey. I appreciate Tarasha Khurana, Neehar Pari, Mengtian Li, Aayush Bansal, Achal Dave, Ravi Teja Mullapudi, Xindi Wu, Gengshan Yang, Haithem Turki, Jason Zhang, Jonothon Luiten, Peiyun Hu, and Kangle Deng for their valuable discussions, constructive feedback, and patience with my occasional abuse of compute clusters. Their support, camaraderie, and shared passion for research have greatly enriched my experience and contributed significantly to my personal and professional growth.

During my time at CMU, I had the privilege of working with a remarkable group of individuals. I would like to express my heartfelt gratitude to all my collaborators who have played a crucial role in my projects. Special thanks to Tiffany Yutong Ling, Shihong Liu, Ryan Lee, Samuel Yu, Zhiyi Kuang, Xinyue Chen, Shubham Parashar, Tian Liu, Jia Shi, Siqi Zeng, Shihao Shen, Jiayao Li, Simran Khanuja, Baiqi Li, Jean de Dieu Nyandwi, Anoushka Shrivastava, Wenxuan Peng, and Prof. Shu Kong. Their unique perspectives, expertise, and dedication have been indispensable to the success and innovation of our collaborative efforts.

Contents

1	Introduction	1
2	Multimodality Helps Unimodality: Cross-Modal Few-Shot Learning with Multimodal Models	3
2.1	Related Works	6
2.2	Cross-Modal Adaptation	9
2.3	Vision-Language Adaptation	12
2.4	Vision-Audio Adaptation	17
2.5	Ablation Studies	18
2.6	Discussion and Limitations	21
3	Language Models as Black-Box Optimizers for Vision-Language Models	25
3.1	Introduction	26
3.2	Related Works	28
3.3	Prompting VLMs Using Chat-Based LLMs	29
3.4	Illustrative Few-Shot Classification Task	31
3.5	More Benefits of Natural Language Prompts	34
3.6	Application: Text-to-Image Generation	35
3.7	Discussion and Limitations	37
4	Revisiting the Role of Language Priors in Vision-Language Models	45
4.1	Introduction	46
4.2	Related Works	49
4.3	The role of language priors	50
4.4	Experimental results on I-to-T retrieval	52
4.5	Additional Experimental Results	55
4.6	Comparison to PMI^k	57
4.7	Ablation Studies on α -Debiasing	58
4.8	Is VisualGPTScore a Biased Estimator?	59
4.9	Discussion and Limitations	60
5	Conclusions	67

List of Figures

2.1	Human perception is internally cross-modal. When we perceive from one modality (such as vision), the same neurons will be triggered in our cerebral cortex as if we are perceiving the object from other modalities (such as language and audio) [48, 128, 135]. This phenomenon grants us a strong ability to learn from a few examples with cross-modal information [96, 128]. In this work, we propose to leverage cross-modality to adapt multimodal models (such as CLIP [154] and AudioCLIP [54]), that encode different modalities to the same representation space.	4
2.2	Adding additional modalities helps few-shot learning. Adding textual labels to a 2-shot cat-vs-dog classification task leads to better test performance (by turning the problem into a 3-shot cross-modal task!). We visualize cross-modal CLIP [45] features (projection to 2D with principal component analysis) and the resulting classifier learned from them, and observe a large shift in the decision boundary. See Figure 2.5 for more examples.	7
2.3	Cross-modality reduces the ambiguity of few-shot learning. Classic (uni-modal) few-shot learning is often <i>underspecified</i> . Even for binary classification, when given only a single image per class (left), it is unclear whether the target class is the animal, the hat, or the background scene. Adding an extra modality, such as text or audio, helps clarify the problem setup (right). Notably, language usually comes “for free” in classification datasets in the form of a textual label per class.	8
2.4	Uni-modal (left) vs. cross-modal adaptation (right). Prior work [45, 195, 214, 221] performs uni-modal adaptation by calculating the loss over a single modality. Cross-modal adaptation makes use of additional training samples from other modalities, exploiting pre-trained encoders that map different modalities to the same representation space. We show that cross-modal learning can also improve prior art and even extends to audio modalities with AudioCLIP [54].	10

2.5	Additional PCA projection plots for random pairs of classes in ImageNet [30]. Adding one-shot text as training samples can oftentimes aggressively shift the decision boundary.	14
3.1	Prompting VLMs using chat-based LLMs. Similar to how human prompt engineers iteratively test and refine prompts, we employ ChatGPT [139, 141] to continuously optimize prompts for vision-language models (VLMs). Our iterative approach assesses the performance of ChatGPT-generated prompts on a few-shot dataset (highlighted in blue) and provides feedback (marked in violet) to ChatGPT through simple conversations, as depicted in the illustrative figure. This straightforward method delivers state-of-the-art results for one-shot image classification across 11 datasets using CLIP, operated in a black-box manner without accessing model weights, feature embeddings, or output logits. We show that providing both positive (in green) and negative prompts (in red) enhances efficiency. Remarkably, our approach outperforms both white-box methods such as gradient-based continuous prompting (CoOp [219]) and human-engineered prompts [153] in this extremely low-shot scenario. This figure only shows a typical conversation using ChatGPT’s web user interface. Our code implementation follows this pattern using the ChatGPT API.	38
3.2	Conversational feedback incorporating both positive and negative prompts leads to improved efficiency. We fix the number of restarts to 20 and iterations to 10, and ablate different numbers of resets on all 11 datasets (left) and ImageNet (right). Notably, our approach using “P+N” (both top-15 and bottom-15 prompts) can optimize faster within a much fewer number of resets than using “P-Only” (top-30 prompts), resulting in the highest overall performance.	40

3.3 **Improving text-to-image (T2I) generation using chat-based *multimodal* LLMs.** We apply our framework to optimize prompts for the state-of-the-art black-box generative VLM, DALL-E 3 [11], using the multimodal GPT4-V [139]. For complicated user queries that DALL-E 3 may initially fail to generate, we send the generated image (in **violet**) along with the current prompt to GPT4-V to ask for feedback on improvements (in **red**) and then generate a new prompt (in **blue**). We show that such a simple framework is surprisingly effective at correcting DALL-E 3 mistakes on some challenging Winoground [182] text queries that involve action, logical, and spatial reasoning. We conduct a human evaluation on the quality of generated images in Table 3.6. We open-source our code at [link](#) to facilitate future research on AI-driven content generation. 41

3.4 **Prompt inversion using chat-based *multimodal* LLMs.** We apply our framework to reverse engineer the text prompt to generate the same user-queried image. We send the generated image (in **violet**) along with the original image to GPT4-V to ask for feedback on improvements (in **red**) and then generate a new prompt (in **blue**). . . 42

4.1 **Two train-test shifts encountered in image-to-text retrieval tasks.** Scenario 1 (**left**) constructs negative captions by shuffling words in the true caption (as in ARO-Flickr), but this produces implausible text such as “white a duck spreads its wings in while the water”. Here, exploiting the language bias of the training set will help since it will downweight the match score for such implausible negative captions. In fact, we show that a blind language-only model can easily identify the correct caption. Scenario 2 (**right**) constructs negative captions that are curated to be plausible (as in SugarCrepe). Here, the language bias of the training set may hurt, since it will prefer to match common captions that score well under the language prior; i.e., the incorrect caption of “people are cooking in a kitchen” is more likely than the true caption of “people are posing in a kitchen” under the language prior, and so removing the language bias improves performance. 47

4.2 **Estimating $P_{train}(\mathbf{t}|\mathbf{i})$ and $P_{train}(\mathbf{t})$ from generative VLMs.** Figure (a) shows how image-conditioned language models such as Li et al. [103] that generate text based on an image can be repurposed for computing $P_{train}(\mathbf{t}|\mathbf{i})$, which is factorized as a product of $\prod_{k=1}^m P(t_k|t_{<k}, \mathbf{i})$ for a sequence of m tokens. These terms can be efficiently computed in *parallel*, unlike *sequential* token-by-token prediction for text generation. Figure (b) shows two approaches for Monte Carlo sampling of $P_{train}(\mathbf{t})$. While the straightforward approach is to sample trainset images, we find that using as few as three “null” (Gaussian noise) images can achieve more robust estimates. 53

List of Tables

2.1	Comparison to SOTA using the CoOp [221] protocol , which reports top-1 accuracy across 11 test sets in Table 2.5. For a fair comparison, we reuse the same few-shot visual samples and hand-engineered text prompts used by Tip-Adapter [214]. The original Tip-Adapter searches over hyperparameters (e.g. early stopping) on the large-scale test set, which may not be realistic for few-shot scenarios. Instead, we rerun their codebase and early-stop on a few-shot validation set (as we do), denoted by †. We reproduce WiSE-FT in our codebase since the original work does not provide few-shot results. In summary, by incorporating one-shot text samples into our training set, a simple cross-modal linear probe already outperforms <i>all</i> prior methods across <i>all</i> shots. Additionally, partial finetuning further improves performance, especially for 8 and 16 shots. Finally, our methods are faster to train than prior work, sometimes significantly (full report in Table 2.8).	15
2.2	Cross-modal adaptation improves existing methods . We follow the same protocol as Table 2.1, reporting the delta accuracy between uni-modal and cross-modal variants of various state-of-the-art methods. The consistent boost suggests that cross-modal training is orthogonal to techniques for uni-modal adaptation, such as prompting [221], adapter [73], and robust finetuning [195].	16
2.3	Image classification results on ImageNet-ESC benchmark . Adding one audio shot can improve image classification under most few-shot scenarios, even when the audio and vision modalities are only loosely aligned.	19
2.4	Audio classification results on ImageNet-ESC benchmark . Similar to Table 2.3, adding one image shot improves few-shot audio classification.	19
2.5	Detailed statistics of the 11 datasets . We adopt the hand-engineered templates selected by Tip-Adapter [214] unless otherwise stated. Note that this set of templates is identical to the ones selected by CLIP [154] and	
2.6	CoOp [221], except for ImageNet Augmentation for cross-modal adaptation . We evaluate the impact of selected augmentation techniques following the same CoOp protocol as in Table 2.1.	20 21

2.7	Robustness under test-time distribution shifts. We follow CoOp [221]’s protocol for evaluating the test-time performance on variants of ImageNet. We report results with two image encoders (ResNet50 and ViT-B/16), and mark the best and <u>second best</u> results. Salient conclusions: (a) Cross-modal linear probing is much more robust than its uni-modal counterpart while being competitive to previous SOTA methods such as WiseFT and CoOp, and (b) it can be further augmented with post-hoc modification through WiseFT to achieve new the SOTA.	22
2.8	Efficiency and accuracy for different methods on ImageNet-16-shot. All experiments are tested with batch size 32 on a single NVIDIA GeForce RTX 3090 GPU. Our approaches take less time and achieve SOTA performance.	23
3.1	Comparison of our method with other baselines on one-shot classification tasks. We report the average accuracy of each method across three folds, optimized using 1-shot training sets. We bold the best black-box result for each dataset, and <u>underline</u> the second best result. First, we note that our approach can effectively improve upon the initial prompts selected from LAIONCOCO-1M from 56% to 61%. Our approach is also competitive against the best Human-Engineered prompts released by OpenAI [153] searched using <i>test set</i> performance. Additionally, we show that using both positive and negative prompts improves the overall accuracy by 1%. For reference, we report <i>oracle</i> white-box approaches in gray. Remarkably, we also surpass white-box solutions such as WiSE-FT [194] and CoOp [219] by 1.5%. These methods require either gradient-based fine-tuning (CoOp/WiSE-FT/Cross-Modal) or prompt ensembling using output logits (DCLIP). While our approach is less effective than the SOTA white-box method (Cross-Modal Adaptation), we stress that our black-box setup is significantly more challenging, because we restrict the optimization space to <i>natural language</i> and do <i>not</i> access the pre-trained weights, model architectures, feature embeddings, and output logits of VLMs.	39
3.2	Example templates returned by our algorithm on each dataset. Although we do not provide ChatGPT with any information regarding the targeted dataset, we observe that the resulting templates are remarkably similar to human-engineered templates, with many domain-specific details such as “motion” and “cuisine”, and stylistic elements such as “bright and natural lighting”.	40

3.3	Black-box prompt transfer from ResNet-50 to other CLIP architectures. We evaluate both our natural language prompts and CoOp’s continuous prompts on 16-shot ImageNet, which are trained using the RN50 CLIP backbone. As a reference point, we include the baseline prompt “a photo of a {}”, and show that the prompts derived from our method using RN50 consistently surpass it after transferring to different backbones. In contrast, while CoOp achieves better 16-shot ImageNet performance using RN50, its performance plummets during the transfer, e.g., from 63% to a mere 21% for RN101.	41
3.4	Examples of T2I optimization. We show that our framework (Figure 3.3) can automatically improve the faithfulness of images generated by DALL-E 3, with respect to user-specified textual topics (for T2I generation) or reference images (for prompt inversion). This is achieved through three rounds of prompt optimization, using feedback from the multimodal LLM (GPT4-V).	42
3.5	Customization via prompt inversion. Users can simply append extra descriptions to the inverted prompts to customize their main characters in queried images.	43
3.6	Our method enhances faithfulness in T2I generation. We hire two human annotators to assess the faithfulness of images generated from user queries, e.g., textual topics for Text-to-Image, or reference images for Prompt Inversion. The scores are measured on a 1-to-5 Likert scale, with 1 signifying contradiction and 5 indicating perfect alignment with the user’s goal. Our approach benefits from three iterations of prompt optimization and consistently outperforms human-engineered prompts by designers who have one year of experience in AI content generation.	43
4.1	OTS generative VLMs are SOTA on image-to-text retrieval benchmarks. We begin by evaluating blind language models (in red). Surprisingly, this already produces SOTA accuracy on certain benchmarks such as ARO-Flickr, compared to the best discriminative approaches (in gray). We also find that blind inference of generative VLMs, $P_{train}(\mathbf{t})$ via sampling Gaussian noise images (in blue), often performs better and achieve above-chance performance even on the most recent SugarCrepe. Next, we show that simply repurposing a generative VLM’s language generation head for computing image-text scores (VisualGPTScore in yellow), which corresponds to $\alpha = 0$, consistently produces SOTA accuracy across all benchmarks. Finally, debiasing this score by tuning α on val set (in green) further improves performance, establishing the new SOTA.	62

4.2	α-debiasing on I-to-T benchmarks and $P_{train}(\mathbf{t})$ frequency charts of both positive and negative captions. Increasing α from 0 to 1 hurts performance on benchmarks with non-sensical negative captions such as ARO and Crepe. Such negative captions are easier to identify because of their low score under the language prior $P_{train}(\mathbf{t})$, implying such benchmarks may even be solved with blind algorithms that avoid looking at images. On the other hand, for benchmarks like SugarCrepe with more balanced $P_{train}(\mathbf{t})$ between positives and negatives, tuning α may lead to performance gain.	63
4.3	Additional results on Winoground/EqBen/COCO/Flickr30K/ImageNet1K. Table (a) shows the importance of α -debiasing on these compositionality and large-scale retrieval benchmarks. While OTS generative scores do not work well, debiasing with a larger α close to 1 can consistently and often significantly improve I-to-T performance. To highlight the improvement, we mark results without debiasing ($\alpha = 0$) (in yellow), debiasing with a fixed $\alpha = 1$ (in pink), and cross-validation using held-out val sets ($\alpha = \alpha_{val}^*$) (in green). Table (b) shows that OTS generative scores can obtain favorable results on all T-to-I retrieval tasks, competitive with the ITMScore.	64
4.4	α-debiasing consistently improves BLIP-2 on balanced VL benchmarks. We show that α -debiasing, even with a fixed $\alpha=1$, can consistently improve BLIP-2 performance on challenging Winoground and EqBen.	64
4.5	Comparing sampling of Gaussian noise images and trainset images for estimating $P_{train}(\mathbf{t})$. We report text scores of α -debiasing on Winoground I-to-T retrieval task. We ablate 3/10/100/1000 Gaussian noise and LAION samples and report both mean and std using 5 sampling seeds. The optimal $\alpha^* \in [0, 1]$ is searched on testset via a step size of 0.001. The Gaussian noise images are sampled with a mean calculated from the LAION subset and a fixed std of 0.25.	64
4.6	I-to-T retrieval on COCO/Flickr30k using different sampling methods. Estimating $P_{train}(\mathbf{t})$ by averaging the scores of testset images (with zero computational cost) demonstrates superior performance compared to sampling additional Gaussian noise images.	65
4.7	α-debiasing results on both val set and test set for COCO/Flickr30k I-to-T retrieval. We observe that validation and test performance are strongly correlated while we interpolate $\alpha \in [0, 1]$	65

4.8	Retrieval performance on randomly sampled LAION114M subsets with varied sizes. Table (a) shows that while OTS generative scores are robust for T-to-I retrieval, its performance degrades on I-to-T retrieval tasks when the number of candidate texts increases. This implies that OTS generative scores suffer from language biases towards certain texts even in the training set. Nonetheless, we show that our debiasing solution using either $\alpha = 1$ or optimal $\alpha^* \in [0, 1]$ with a step size of 0.001, can consistently boost the performance. Figure (b) visualizes α -debiasing results on LAION subsets, where each curve represents a different sample size.	65
-----	--	----

Chapter 1

Introduction

The evolution of vision-language foundation models has been pivotal not only in academic research but also in transforming everyday tasks. Models like CLIP have revolutionized visual recognition, becoming foundational in a variety of multimodal systems and downstream applications. GPT4-Vision, with its advanced visual understanding, is adept at tasks such as chart reading and GUI navigation and can already serve as a professional assistant. Similarly, models like DALL-E 3 stand at the forefront of creative generation, turning text prompts into designer-quality images. This thesis explores innovative methods to enhance the functionality of vision-language models (VLMs), while also integrating insights from large language models (LLMs) like ChatGPT to optimize their performance in a wider array of applications.

The first approach, termed "cross-modal adaptation," is a significant breakthrough in aligning discriminatively-pretrained VLMs with user objectives, particularly in visual classification tasks. This method leverages textual or audio data to construct better visual classifiers with minimal supervision, surpassing concurrent prompting methods in efficiency and efficacy. It represents a paradigm shift in using multiple modalities to enhance unimodal recognition capabilities.

Building on this, the second approach introduces a truly black-box method that employs LLMs, such as ChatGPT, as natural language prompt optimizers for VLMs. This approach demonstrates superior results in one-shot visual classification and text-to-image optimization tasks, including image generation and prompt inversion.

1. Introduction

By harnessing the strengths of both VLMs and LLMs, this method showcases the potential of integrated AI systems in complex applications.

While these adaptation methods have shown effectiveness, they do not directly address the inherent limitations of current vision-language foundation models in dealing with compositional and complex reasoning. The final part of this thesis focuses on these fundamental challenges. It presents strategies to assess and enhance the models' capabilities in handling detailed compositions of objects, attributes, and their relationships. This is crucial for applications requiring fine-grained control and sophisticated reasoning, pushing the limits of what vision-language models can achieve and better aligning them with nuanced user goals.

Chapter 2

Multimodality Helps Unimodality: Cross-Modal Few-Shot Learning with Multimodal Models

The ability to quickly learn a new task with minimal instruction – known as few-shot learning – is a central aspect of intelligent agents. Classical few-shot benchmarks make use of few-shot samples from a single modality, but such samples may not be sufficient to characterize an entire concept class. In contrast, humans use cross-modal information to learn new concepts efficiently. In this work, we demonstrate that one can indeed build a better **visual** dog classifier by **reading** about dogs and **listening** to them bark. To do so, we exploit the fact that recent multimodal foundation models such as CLIP are inherently cross-modal, mapping different modalities to the same representation space. Specifically, we propose a simple **cross-modal adaptation** approach that learns from few-shot examples spanning different modalities. By repurposing class names as additional one-shot training samples, we achieve SOTA results with an embarrassingly simple linear classifier for vision-language adaptation. Furthermore, we show that our approach can benefit existing methods such as prefix tuning, adapters, and classifier ensembling. Finally, to explore other modalities beyond vision and language, we construct the first (to our knowledge) audiovisual few-shot benchmark and use cross-modal training to improve the performance of both image and audio classification.

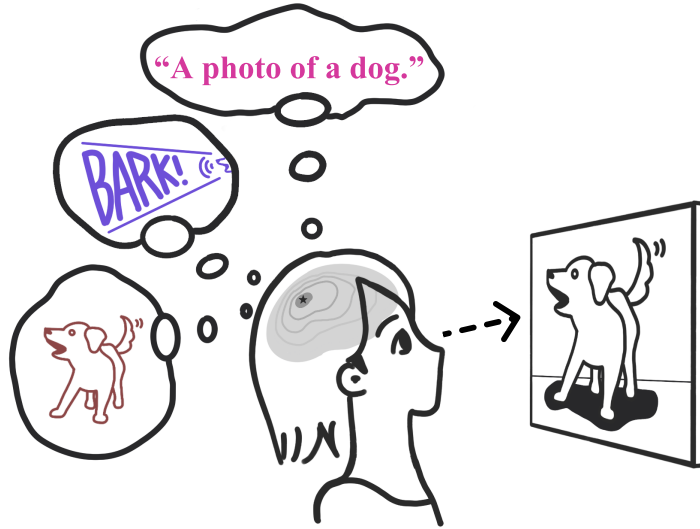


Figure 2.1: **Human perception is internally cross-modal.** When we perceive from one modality (such as vision), the same neurons will be triggered in our cerebral cortex as if we are perceiving the object from other modalities (such as language and audio) [48, 128, 135]. This phenomenon grants us a strong ability to learn from a few examples with cross-modal information [96, 128]. In this work, we propose to leverage cross-modality to adapt multimodal models (such as CLIP [154] and AudioCLIP [54]), that encode different modalities to the same representation space.

Learning with minimal instruction is a hallmark of human intelligence [163, 176, 190], and is often studied under the guise of few-shot learning. In the context of few-shot visual classification [34, 42, 56, 86, 150, 156], a classifier is first pre-trained on a set of base classes to learn a good feature representation and then adapted or finetuned on a small amount of novel class data. However, such few-shot setups often face an inherent ambiguity – if the training image contains a golden retriever wearing a hat, how does the learner know if the task is to find **dogs**, **golden retrievers**, or even **hats**? On the other hand, humans have little trouble understanding and even generalizing from as few as one example. How so?

We argue that humans make use of multimodal signals and representations (Figure 2.1) when learning concepts. For example, verbal language has been shown to help toddlers better recognize visual objects given just a few examples [81, 175]. Indeed, there exists ample evidence from neuroscience suggesting that cognitive

representations are inherently multimodal. For instance, visual images of a person evoke the same neurons as the textual strings of the person’s name [151] and even audio clips of that person talking [135]. Even for infants as young as 1-5 months old, there is a strong correspondence between auditory-visual [96] as well as visual-tactile signals [128]. Such *cross-modal* or inter-modal representations are fundamental to the human perceptual-cognitive system, allowing us to understand new concepts even with few examples [48].

Cross-modal adaptation (our approach). In this paper, we demonstrate that cross-modal understanding of different modalities (such as image-text or image-audio) can improve the performance of individual modalities. That is, *reading* about dogs and *listening* to them bark can help build a better *visual* classifier for them! To do so, we present a remarkably simple strategy for cross-modal few-shot adaptation: *we treat examples from different modalities as additional few-shot examples*. For example, given the “1-shot” task of learning a dog classifier, we treat *both* the textual dog label and the single visual image as training examples for learning a (visual) dog classifier. Learning is straightforward when using frozen textual and visual encoders, such as CLIP [154], that map different modalities to the same representational space. In essence, we have converted the “n-shot” problem to a “(n+1)-shot” problem (Figure 2.2)! We demonstrate that this basic strategy produces SOTA results across the board with a simple linear classifier, and can be applied to existing finetuning methods [195, 214, 221] or additional modalities (e.g. audio).

Why does it work? From one perspective, it may not be surprising that cross-modal adaptation improves accuracy, since it takes advantage of additional training examples that are “hidden” in the problem definition, e.g. a label name [201] or an annotation policy [133] for each class. However, our experiments demonstrate that multimodal cues are often complementary since they capture different aspects of the underlying concept; a dog label paired with a single visual example is often more performant than two images! For example, Figure 2.3 demonstrates a one-shot example where the target concept is ambiguous, but becomes clear once we add information from other modalities like language and sound.

Multimodal adaptation (prior art). In contrast to our cross-modal approach, most prior works simply follow the popular practice of finetuning uni-modal foundation models, such as large vision [23, 59, 60] or language models [16, 33, 118]. For example,

CoOp [221] and other prompting methods [121, 220, 223] finetune CLIP via prefix tuning to replace hand-engineered prompts such as "a photo of a {cls}" with learned word tokens. Similarly, inspired by parameter-efficient tuning of language models [73], adapter-based methods [45, 214] finetune CLIP by inserting lightweight multi-layer-perceptrons (MLPs). However, we aim to study the fundamental question of how to finetune *multi*-modal (as opposed to *uni*-modal) models. A crucial difference between prior art and ours is the use of textual information, as all existing methods [80, 195, 214, 221] repurpose additional text features as *classifier weights* instead of *training samples*. We demonstrate in this paper that cross-modal adaptation is not only more performant but can also benefit prior uni-modal approaches.

Problem setup. We begin by replicating the existing evaluation protocol of other works [154, 214, 221] on few-shot adaptation of vision-language models, and report performance on 11 diverse downstream datasets. We produce state-of-the-art accuracy with an embarrassingly simple linear classifier that has access to additional "hidden" training examples in the form of textual labels, resulting in a system that is far more lightweight than prior art. Interestingly, we show that existing approaches [195, 214, 221], despite already repurposing text features as classifier weights, can still benefit from cross-modal learning. Finally, we extend our work to the audio domain by taking advantage of AudioCLIP [54] that maps audio to the same frozen CLIP representation space. We construct the first (to our knowledge) *cross-modal few-shot learning benchmark with audio* by intersecting ImageNet [30] and the ESC-50 audio classification dataset [147]. We show that cross-modal audiovisual learning helps for both downstream image and audio classification; in summary, one *can* train better dog image classifiers by listening to them bark!

2.1 Related Works

Webly-supervised pre-training. Learning *foundation models* [12] from large-scale web data is becoming a predominant paradigm in AI. In NLP, models such as BERT [33] and GPT-3 [16] are pre-trained on a massive web text corpus with language-modeling objectives and can be transferred to a wide range of downstream tasks, even without explicit supervised finetuning [115, 183]. Self-supervision [20, 23, 59] is also a trending topic in the vision community, and recent methods [50, 60] demonstrate

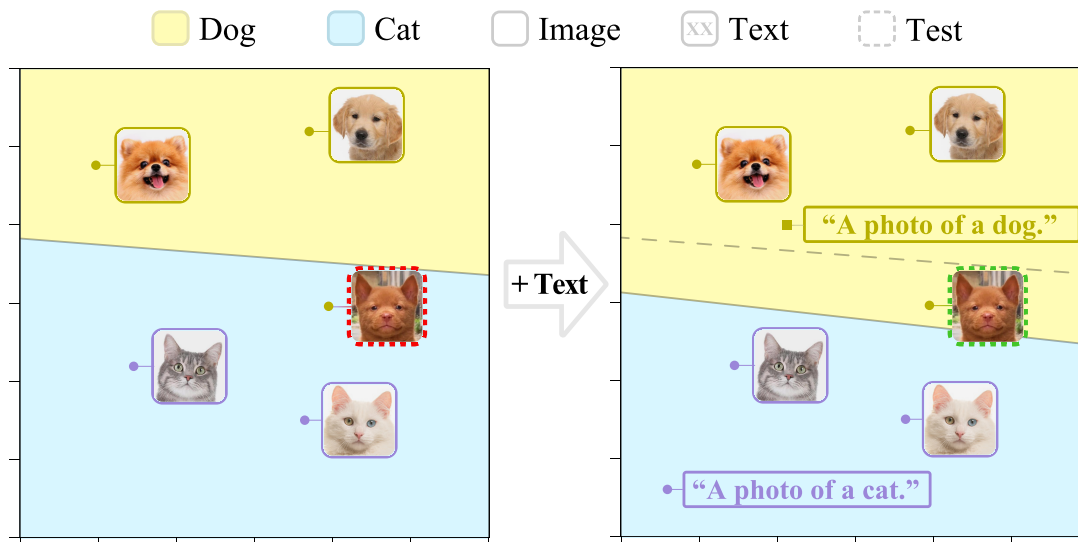


Figure 2.2: **Adding additional modalities helps few-shot learning.** Adding textual labels to a 2-shot cat-vs-dog classification task leads to better test performance (by turning the problem into a 3-shot cross-modal task!). We visualize cross-modal CLIP [45] features (projection to 2D with principal component analysis) and the resulting classifier learned from them, and observe a large shift in the decision boundary. See Figure 2.5 for more examples.

even stronger visual representations than fully-supervised pre-trained ones such as on ImageNet [30].

Multimodal foundation models. Recently, foundation models have shifted towards a multimodal supervision paradigm. For visual representation learning, early works transform web image captions into structured outputs for supervised learning, such as multi-label targets [87] or visual n-grams [99]. More recently, CLIP [154] and ALIGN [83] propose a simple contrastive-based approach to embed images and captions into the same representation space, and demonstrate impressive “zero-shot” performance on downstream tasks. Follow-up works enhance multimodal pre-training by incorporating generative-based objectives [4, 104, 206], consistency regularization [111, 134], stronger visual priors [210], phrase-grounding tasks [106, 212], and audiovisual information through videos [54]. In this work, we focus on adapting CLIP [154] and AudioCLIP [54] for few-shot classification because contrastive-based multimodal models are stronger classifiers [4]. Adopting other multimodal models [4,

2. Multimodality Helps Unimodality: Cross-Modal Few-Shot Learning with Multimodal Models

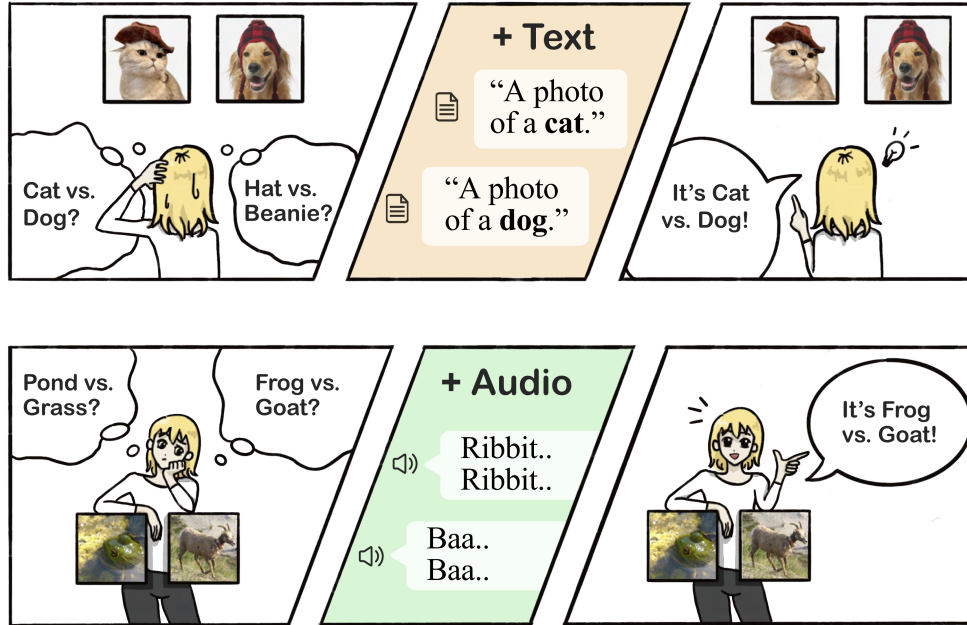


Figure 2.3: **Cross-modality reduces the ambiguity of few-shot learning.** Classic (uni-modal) few-shot learning is often *underspecified*. Even for binary classification, when given only a single image per class (**left**), it is unclear whether the target class is the animal, the hat, or the background scene. Adding an extra modality, such as text or audio, helps clarify the problem setup (**right**). Notably, language usually comes “for free” in classification datasets in the form of a textual label per class.

206] or adapting to tasks other than classification [177, 212] can be interesting future directions.

Adaptation of foundation models. As multimodal pre-trained models have excelled at classic vision tasks [154, 212], there has been surging interest in developing more efficient adaptation methods. However, we observe that most of the trending techniques are built upon successful recipes crafted for uni-modal foundation models. For example, CLIP [154] adopts linear probing [23, 59, 60, 212] and full-finetuning [49, 60, 91, 189, 196, 212] when transferring to downstream tasks. Prompt adaptation of CLIP [121, 154, 202, 220, 223] is motivated by the success of prefix-tuning for language models [31, 46, 57, 85, 115, 148, 161, 162, 171]. Similarly, CLIP-Adapter [45] and Tip-Adapter [214] are inspired by parameter-efficient finetuning methods [73, 84, 213] that optimize lightweight MLPs while freezing the encoder. Yet, all aforementioned methods including WiSE-FT [195] use the other modality, e.g. textual labels, as

classifier weights and still calculate a *uni-modal* softmax loss on the few-shot images. We instead show that incorporating other modalities as *training samples* is far more effective.

Few-shot classification. Prior successful few-shot learning methods leverage meta learning [42, 156], metric learning [8, 176, 184], transfer learning [56, 150], and transductive learning [34, 86]. These classic algorithms usually assume a large meta-training set for pre-training the network, and then evaluate on multiple episodes of few-shot train (support) and test (query) sets. In this work, we instead follow the new evaluation protocol implemented by recent works on few-shot adaptation with CLIP [154, 214, 221]: (1) the meta-training phase is replaced with pre-trained CLIP models, and (2) the test sets are the official test splits of each dataset (thus not few-shot). Notably, none of the prior works [214, 221] we compare to in this paper perform optimization with test set samples, and we follow this practice to ensure a fair comparison. We leave semi-supervised [188] or transductive finetuning [34, 77] techniques as future work.

Cross-modal machine learning. Inspired by cross-modal human cognition [18, 93, 135], cross-modal learning [133, 201] is a subfield of multimodal machine learning [1, 5, 19, 70, 97, 109, 122, 142, 143, 167, 211] that aims to use data from additional modalities to improve a uni-modal task. Cross-modal learning does not require instance-wise alignment; for example, existing algorithms [133, 201] can benefit from class-level descriptions as opposed to image-level captions. In this work, we propose a lightweight cross-modal learning method by treating data from other modalities as additional training samples. Furthermore, we encourage future works to embrace cross-modal few-shot learning as opposed to the underspecified uni-modal setup (Figure 2.3).

2.2 Cross-Modal Adaptation

In this section, we mathematically formalize our approach to cross-modal few-shot learning.

Uni-modal learning. We begin by reviewing standard uni-modal few-shot classification, which learns a classifier from a small dataset of (x_i, y_i) pairs and

2. Multimodality Helps Unimodality: Cross-Modal Few-Shot Learning with Multimodal Models

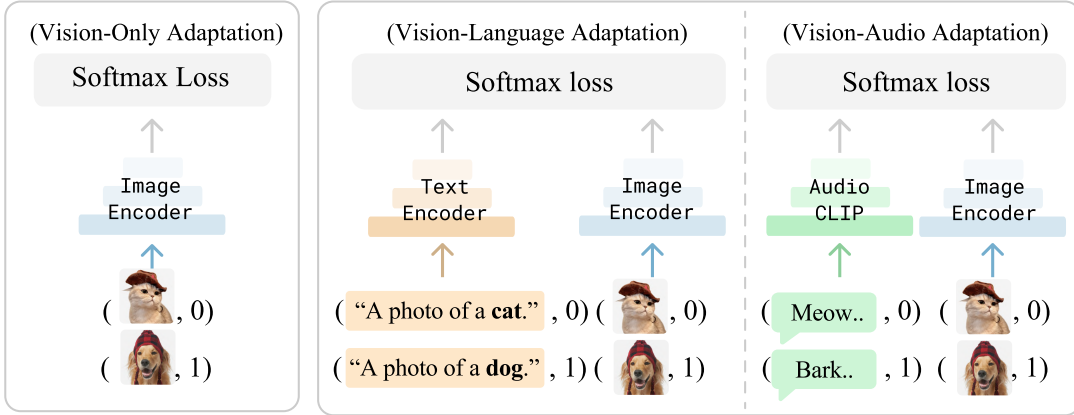


Figure 2.4: **Uni-modal (left) vs. cross-modal adaptation (right)**. Prior work [45, 195, 214, 221] performs uni-modal adaptation by calculating the loss over a single modality. Cross-modal adaptation makes use of additional training samples from other modalities, exploiting pre-trained encoders that map different modalities to the same representation space. We show that cross-modal learning can also improve prior art and even extends to audio modalities with AudioCLIP [54].

pre-trained feature encoder $\phi(\cdot)$:

$$\mathcal{L}_{uni-modal} = \sum_i \mathcal{H}(y_i, \phi(x_i)) \quad (2.1)$$

where \mathcal{H} is typically the softmax loss

$$\mathcal{H}(y, f) = -\log(p(y|f)) = -\log\left(\frac{e^{w_y \cdot f}}{\sum_{y'} e^{w_{y'} \cdot f}}\right). \quad (2.2)$$

Our notation separates the feature extractor ϕ from the final class weights w_y , since the former is typically pre-trained on a massive source dataset and the latter is trained on the few-shot target dataset. However, sometimes the representation ϕ can also be finetuned on the few-shot dataset (as we explore in our experiments). Importantly, both the class weights and feature extractor must live in the same N -dimensional space in order to compute their inner product:

$$w_y, \phi(\cdot) \in \mathbb{R}^N. \quad (2.3)$$

Though we focus on classification, class models could be learned via other losses (such as centroid prototypes [176]).

Cross-modal learning. Our extension to multiple modalities is straightforward; we assume each training example is accompanied by a discrete label m denoting its modality:

$$(x_i, y_i) \rightarrow (x_i, y_i, m_i), \quad x_i \in X_{m_i}, \quad m_i \in M. \quad (2.4)$$

For example, one may define the set of modalities to be $M = \{\text{visual, language}\}$ or $\{\text{visual, audio}\}$ (Figure 2.4). We can then define an associated loss:

$$\mathcal{L}_{\text{cross-modal}} = \sum_i \mathcal{H}(y_i, \phi_{m_i}(x_i)), \quad (2.5)$$

where we crucially assume access to modality-specific feature encoders ϕ_m for $m \in M$. While the individual datapoints x_i may come from different modalities with different dimensions, our formulation requires that the encoders map all modalities to the same fixed-dimensional space.

$$w_y, \phi_m(\cdot) \in R^N. \quad (2.6)$$

Note that this requirement is satisfied by many multimodal foundation models such as CLIP [154] and ALIGN [83] since they map different modalities into the same N -dimensional embedding.

Inference: The learned classifier can produce a label prediction for a test example x from *any* modality $m \in M$:

$$\hat{y} = \operatorname{argmax}_{y'} w_{y'} \cdot \phi_m(x). \quad (2.7)$$

This means we can use the same classifier to classify different test modalities (e.g. images and audio clips). In this paper, we mainly evaluate on a single modality (like images) to emphasize that *multimodality helps unimodality*.

Cross-modal ensembles. We now show that cross-modal learning produces classifiers that are ensembles of modality-specific classifiers, exposing a connection to

related approaches for ensembling (such as WiSE-FT [195]). We begin by appealing to the well-known *Representer Theorem* [164], which shows that optimally-trained classifiers can be represented as linear combinations of their training samples. In the case of a cross-modal linear probe, weights for class y must be a weighted combination of all i training features, across all modalities:

$$w_y = \sum_i \alpha_{iy} \phi_{m_i}(x_i) = \sum_{m \in M} w_y^m, \quad \text{where}$$

$$w_y^m = \sum_{\{i:m_i=m\}} \alpha_{iy} \phi_m(x_i). \quad (2.8)$$

Linear classification via cross-modal adaptation solves for all weights α_{iy} *jointly*, so as to minimize the empirical risk (or training loss). In contrast, prior art optimizes for image-specific α_{iy} ’s *independently* of the text-specific α_{iy} ’s, linearly combining them with a single global α (as in WiSE-FT [195]) or via text-based classifier initialization [45, 214]. Our analysis suggests that the joint optimization enabled by cross-modal learning may help other adaptation methods, as our experiments do in fact show.

Extensions. Although we focus on uni-modal inference tasks (e.g. image classification), the above formulation allows the learned classifier to be applied to *multimodal* test sets, such as classifying videos by training on image and audio, and then ensembling predictions across the two modalities with [Equation 2.7](#). Or, one can extend image classification by providing additional data such as captions and/or attributes. We leave these scenarios as future work. Finally, just as one can optimize uni-modal losses (2.1) by finetuning the encoder ϕ , one can similarly finetune modality-specific encoders ϕ_m in the cross-modal setting (2.5). We explore this finetuning method in the next section.

2.3 Vision-Language Adaptation

We now explore our cross-modal formulation for a particular multimodal setting. Many prior works [133, 201, 214, 221] explore the intersection of vision and language, and thus that is our initial focus. Interestingly, the influential “zero-shot” and “few-shot” evaluation protocols introduced by prior work [154, 198] can be mapped to our

cross-modal setting, with one crucial difference; the textual label of each class can be treated as an explicit training sample (x_i, y_i, m_i) . From this perspective, “zero-shot” learning may be more naturally thought of as one-shot cross-modal learning that learns a few-shot model on *text* and then infers with it on *images*.

Few-shot evaluation protocol. To ensure a fair comparison, we strictly follow the protocol of CoOp [221] by reporting test performance on 11 public image datasets (Table 2.5), with ResNet50 [58] as the image encoder backbone. For maximal reproducibility, we use CoOp’s dataset splits [221] and the three-fold few-shot train sets sampled with the same random seeds. We adopt the given test split of each dataset as the test set. Some prior works [121, 214] apparently use the large-scale test set to tune hyperparameters for few-shot learning; we instead exercise due diligence by tuning hyperparameters (such as the learning rate, weight decay, and early stopping) on the given few-shot validation set with $\min(n, 4)$ examples, where n is the number of training shots.

Cross-modal adaptation outperforms SOTA. Table 2.1 shows the effectiveness of our proposal: we surpass all prior art with an embarrassingly simple linear classifier that requires significantly less training time than other carefully-crafted algorithms. In addition, partial finetuning of the last attentional pooling layer from ϕ_{image} sets the new SOTA. To ensure a fair comparison, we augment the class names into sentences using hand-engineered templates selected by Tip-Adapter [214] (Table 2.5) and follow their practice to initialize the linear layer with text features. Furthermore, we perform minimal image augmentation with a center crop plus a flipped view instead of random crops as in prior art [214, 221]. As such, we can pre-extract features before training the classifier, leading to significantly less training time as shown in Table 2.8. We also show that our method can benefit from both image and text augmentation in Table 2.6.

Why does cross-modal learning help? As stated earlier, one reason that cross-modal learning helps is that it turns the original n -shot problem to an $(n + 1)$ -shot one. However, Table 2.1 shows that 1-shot cross-modal linear probing outperforms the 2-shot results of most prior methods. This suggests that training samples from other modalities tend to contain complementary cues [133, 195, 201]. One can loosely observe this in Figure 2.2 and Figure 2.5, whereby visual and text examples lie in slightly different parts of the embedding space (indicating the potential to aggressively

2. Multimodality Helps Unimodality: Cross-Modal Few-Shot Learning with Multimodal Models

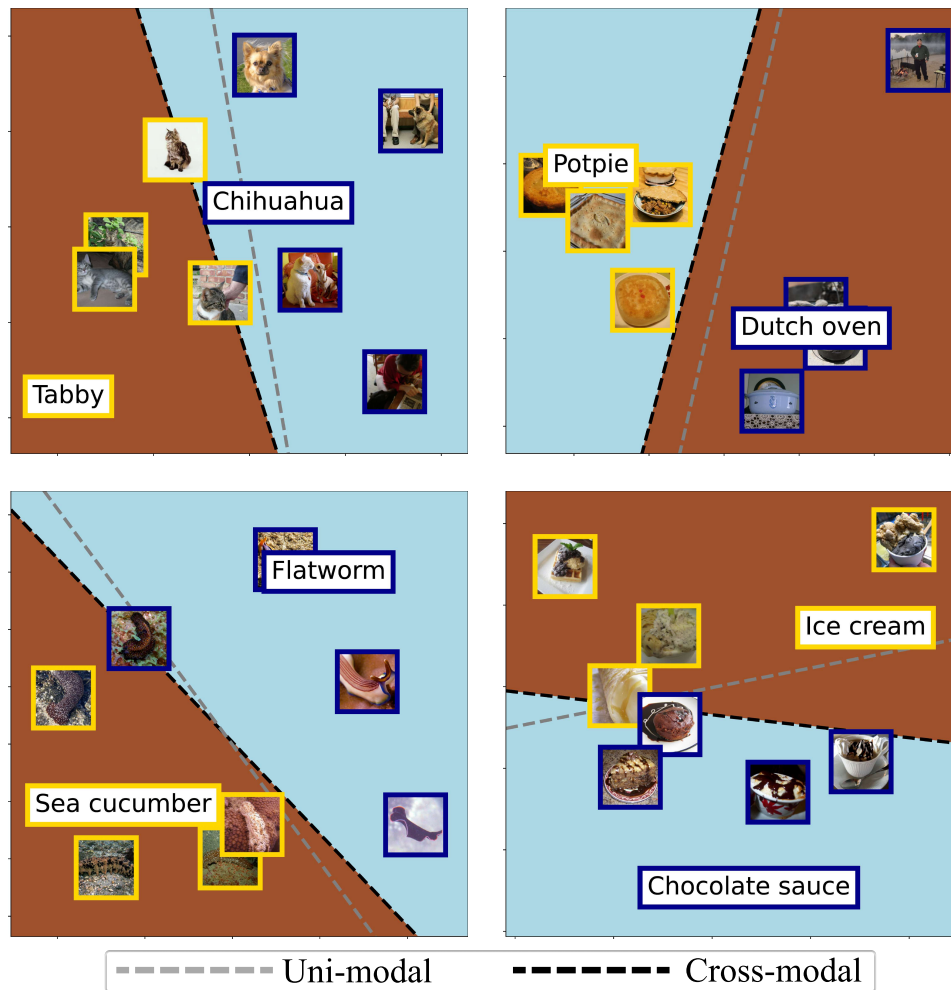


Figure 2.5: **Additional PCA projection plots for random pairs of classes in ImageNet [30].** Adding one-shot text as training samples can oftentimes aggressively shift the decision boundary.

Method	Number of shots					Train speed
	1	2	4	8	16	
Zero-Shot CLIP (58.8)	-	-	-	-	-	-
Linear Probing	36.7	47.6	57.2	65.0	71.1	<1min
WiSE-FT [195]	59.1	61.8	65.3	68.4	71.6	<1min
CoOp [221]	59.6	62.3	66.8	69.9	73.4	14hr
ProGrad [223]	62.6	64.9	68.5	71.4	74.0	17hr
Tip-Adapter [214]	64.5	66.7	69.7	72.5	75.8	5min
Tip-Adapter [†] [214]	63.3	65.9	69.0	72.2	75.1	5min
Cross-Modal Linear Probing	64.1	67.0	70.3	73.0	76.0	<1min
Cross-Modal Partial Finetuning	64.7	67.2	70.5	73.6	77.1	<3min

Table 2.1: **Comparison to SOTA using the CoOp [221] protocol**, which reports top-1 accuracy across 11 test sets in Table 2.5. For a fair comparison, we reuse the same few-shot visual samples and hand-engineered text prompts used by Tip-Adapter [214]. The original Tip-Adapter searches over hyperparameters (e.g. early stopping) on the large-scale test set, which may not be realistic for few-shot scenarios. Instead, we rerun their `codebase` and early-stop on a few-shot validation set (as we do), denoted by †. We reproduce WiSE-FT in our codebase since the original work does not provide few-shot results. In summary, by incorporating one-shot text samples into our training set, a simple cross-modal linear probe already outperforms *all* prior methods across *all* shots. Additionally, partial finetuning further improves performance, especially for 8 and 16 shots. Finally, our methods are faster to train than prior work, sometimes significantly (full report in Table 2.8).

shape the final decision boundary). In fact, WiSE-FT [195] is inspired by similar reasons to ensemble the uni-modal visual classifier with a “zero-shot” (one-shot-text) classifier (in the linear probing case). However, Equation 2.8 shows that cross-modal adaptation can also be seen as jointly learning an ensemble, while WiSE-FT [195] learns the visual classifier independently of the text classifier. This suggests that other adaptation methods may benefit from cross-modal learning, as we show next.

Cross-modal adaptation helps prior art (Table 2.2). This includes prompting (CoOp [221]), adapters (Tip-Adapter [214]), and robust-finetuning (WiSE-FT [195]). We see a large improvement in the low-data regime (1 and 2 shots). Notably, we do not need to tune any methods, and simply reuse the reported hyperparameters. For prompting, we follow CoOp [221] to optimize 16 continuous tokens with the same training setting. For the Adapter model, we follow the same 2-layer MLP architecture

Method	Number of shots				
	1	2	4	8	16
Linear Probing	36.7	47.6	57.2	65.0	71.1
Cross-Modal Linear Probing	64.1	67.0	70.3	73.0	76.0
Δ	27.4	19.4	13.1	8.0	4.9
WiSE-FT [195]	59.1	61.8	65.3	68.4	71.6
Cross-Modal WiSE-FT	63.8	66.4	69.0	71.7	74.1
Δ	4.7	4.6	3.7	3.3	2.5
CoOp [221]	59.6	62.3	66.8	69.9	73.4
Cross-Modal Prompting	62.0	64.9	68.6	71.4	74.0
Δ	2.4	2.6	1.8	1.5	0.6
Tip-Adapter [†] [214]	63.3	65.9	69.0	72.2	75.1
Cross-Modal Adapter	64.4	67.6	70.8	73.4	75.9
Δ	1.1	1.7	1.8	1.2	0.8

Table 2.2: **Cross-modal adaptation improves existing methods.** We follow the same protocol as Table 2.1, reporting the delta accuracy between uni-modal and cross-modal variants of various state-of-the-art methods. The consistent boost suggests that cross-modal training is orthogonal to techniques for uni-modal adaptation, such as prompting [221], adapter [73], and robust finetuning [195].

of CLIP-Adapter [45] with the given residual ratio of 0.2; we outperform Tip-Adapter without relying on their training-free initialization of MLP. For WiSE-FT, we adopt the given ratio (0.5) to post-hoc ensemble the learned and the zero-shot classifiers. Overall, our experiments suggest that cross-modal adaptation is consistently effective, and should likely be a baseline moving forward given its ease-of-implementation. For example, instead of separately benchmarking on “zero-shot” (one-shot-text) and few-shot-vision, a cross-modal linear prob would suffice to evaluate representations of a multimodal model.

2.4 Vision-Audio Adaptation

We now explore cross-modal adaption for other modalities such as audio. We pose the following question: can one learn a better dog *visual* classifier by *listening* to a dog barking? To examine this question, we curate the first audiovisual benchmark that supports few-shot classification of both image and audio.

Our ImageNet-ESC benchmark.¹ We construct our audiovisual benchmark by intersecting two of the most popular image and audio datasets: ImageNet [30] with 1000 types of objects and ESC-50 [147] with 50 types of environmental sounds (including animal, nature, human activity, domestic, and urban noises). We use the class names of the two datasets for class matching. For each class in ESC-50, we check whether there is a corresponding ImageNet class that may produce this type of sound. In this process, we observe that the audio-to-object matching can sometimes be one-to-many. For example, the `clock-alarm` class in ESC-50 can be mapped to either `digital clock` or `analog clock` in ImageNet; the `dog (barking)` class in ESC-50 can be matched to any of the 120 dog species. In such scenarios, we randomly match the classes, e.g. `clock alarm` to `digital clock` and `dog` to `otterhound`. Also, we find that some audio classes loosely match with some visual objects, such as `drinking-sipping` to `water bottle` and `pouring-water` to `water jug`. As such, we create two versions of the dataset: (1) **ImageNet-ESC-27**, which represents the *maximal* intersection consisting of all loose matches, and (2) **ImageNet-ESC-19**, a subset of the former version consisting of more accurate matches.

Few-shot evaluation protocol. We use five-fold few-shot splits sampled from ImageNet, with each split divided into half for training and validation. Test performance is recorded on the official ImageNet validation set of the corresponding classes. We adopt the predefined five folds of ESC-50, where each fold contains 8 samples per class. We construct 5 splits from ESC-50 by selecting one fold for training and validation, and record test performance on the other 4 folds. We report averaged performance over 25 runs (since we have 5 random splits for each modality). To keep consistent with our vision-language experiments, we adopt a uni-modal validation and test set and leave cross-modal testing for future work.

¹Download instructions can be found in our [codebase](#).

Audio encoding. We use AudioCLIP [54] with an ESResNeXT backbone [55] as the audio encoder ϕ_{audio} . Because AudioCLIP is trained on a large-scale video dataset (AudioSet [47]) while freezing the pre-trained CLIP text and image encoder, it produces audio embeddings in the same representation space. While AudioCLIP is pretrained on a sizable amount of data, we note that it does not come close to matching the scale of CLIP pretraining [54, 154]. Thus, it does not perform favorably compared to the SOTA for downstream “zero-shot” audio (i.e. one-shot text) classification tasks [54]. However, scaling up audio pretraining is orthogonal to our investigation.

Audio improves image classification. Table 2.3 shows that adding a random one-shot-audio improves upon naive image-only linear probing, especially in an extremely low-shot setting. This reaffirms Figure 2.3’s hypothesis that cross-modality can reduce the ambiguity of the uni-modal few-shot setup; in other words, one can learn a better *image* classifier by *listening* to object sounds. One exception is the 4-shot performance on ImageNet-ESC-27, where adding audio does not help. We posit that (1) loosely-matched classes can result in noisier training data, and (2) the audio representations are not as robust due to smaller-scale pretraining. This suggests that cross-modal adaptation is less effective when representations are not aligned well or insufficiently trained. Nevertheless, under most scenarios, cross-modal adaptation helps. For all experiments, we follow an identical procedure to vision-language experiments in section 2.2.

Vision improves audio classification. We additionally evaluate the *reverse* task – whether adding a random one-shot *image* sample for downstream audio classification can improve upon audio-only training. Table 2.4 shows the results, where we see the same favorable trend. This success concludes that our approach is modality-agnostic.

2.5 Ablation Studies

We present a few selected ablation studies in this section.

Data augmentation of text samples. Like most prior works [154, 221], we also find that data augmentation can improve downstream performance during vision-language adaptation (cf. Table 2.1). Notably, since the class names are included

Dataset	Method	Image Classification		
		1-shot	2-shot	4-shot
ImageNet-ESC-19	Image-Only Linear	68.0	75.7	83.1
	Image-Audio Linear	69.3	76.7	83.2
ImageNet-ESC-27	Image-Only Linear	60.1	71.8	79.0
	Image-Audio Linear	60.9	73.3	78.9

Table 2.3: **Image classification results on ImageNet-ESC benchmark.** Adding one audio shot can improve image classification under most few-shot scenarios, even when the audio and vision modalities are only loosely aligned.

Dataset	Method	Audio Classification		
		1-shot	2-shot	4-shot
ImageNet-ESC-19	Audio-Only Linear	31.2	41.1	48.5
	Audio-Image Linear	35.7	45.9	51.6
ImageNet-ESC-27	Audio-Only Linear	28.2	39.0	47.1
	Audio-Image Linear	35.0	43.5	48.5

Table 2.4: **Audio classification results on ImageNet-ESC benchmark.** Similar to [Table 2.3](#), adding one image shot improves few-shot audio classification.

as training samples, one can explore augmentation techniques for text (just as random cropping for images). Besides the fixed template a photo of a {cls} and hand-crafted templates ([Table 2.5](#)), we also try a **template mining** strategy that does not rely on the selected dataset-specific templates. To automatically mine for the templates, we search among a pool of 180 templates for 21 templates with the best zero-shot performance on the few-shot validation set of each dataset. For image augmentation, we perform standard flipping and random cropping. We show a subset of results in [Table 2.6](#), and find that all text augmentation techniques provide a sizable boost in performance. The salient conclusions include (1) the performance gain from image augmentation is saturated after more than two views, and (2) template

2. Multimodality Helps Unimodality: Cross-Modal Few-Shot Learning with Multimodal Models

Dataset	Classes	Train	Val	Test	Hand-crafted Prompt [214]
Caltech101 [41]	100	4,128	1,649	2,465	a photo of a {cls}.
OxfordPets [146]	37	2,944	736	3,669	a photo of a {cls}, a type of pet.
StanfordCars [95]	196	6,509	1,635	8,041	a photo of a {cls}.
Flowers102 [137]	102	4,093	1,633	2,463	a photo of a {cls}, a type of flower.
Food101 [13]	101	50,500	20,200	30,300	a photo of {cls}, a type of food.
FGVCAircraft [126]	100	3,334	3,333	3,333	a photo of a {cls}, a type of aircraft.
SUN397 [199]	397	15,880	3,970	19,850	a photo of a {cls}.
DTD [27]	47	2,820	1,128	1,692	{cls} texture.
EuroSAT [62]	10	13,500	5,400	8,100	a centered satellite photo of {cls}.
UCF101 [178]	101	7,639	1,898	3,783	a photo of a person doing {cls}.
					itap of a {cls}.
					a bad photo of the {cls}.
					a origami {cls}.
					a photo of the large {cls}.
					a {cls} in a video game.
					art of the {cls}.
ImageNet [30]	1000	1.28M	N/A	50,000	a photo of the small {cls}.

Table 2.5: **Detailed statistics of the 11 datasets.** We adopt the hand-engineered templates selected by Tip-Adapter [214] unless otherwise stated. Note that this set of templates is identical to the ones selected by CLIP [154] and CoOp [221], except for ImageNet.

mining can be as competitive as a large number of 36 carefully-tuned prompts. In fact, prompting [115, 121, 221] can be viewed as another *text augmentation* technique under cross-modal adaptation, and we leave this exploration to future work.

Test-time distribution shifts. We examine how robust our approach is against test-time distribution shifts in Table 2.7. Specifically, we follow the CoOp [221] protocol to report the test performance of a classifier trained on the source dataset (16-shot ImageNet) to 4 distribution-shifted target test sets, including ImageNet-V2 [157], ImageNet-Sketch [185], ImageNet-A [64], and ImageNet-R [63]. As shown in Table 2.7, cross-modal adaptation can significantly boost the robustness of image-only linear probing and is competitive against baselines designed to address robustness such as CoCoOp [220] and WiSE-FT [195]. Cross-Modal adaptation also improves upon WiSE-FT [195] and sets the new SOTA. We can conclude that language modality plays an important role in robustness, similar to how humans rely on textual cues for recognition [64].

Efficiency. As shown in Table 2.8, our approaches are much more lightweight because we do not rely on deep finetuning [220, 221] or heavy image augmentations.

Finetuning	ImageAugment	TextAugment	Number of shots				
			1	2	4	8	16
Linear	CenterCrop	Classname	61.8	65.3	69.0	72.0	74.9
		a photo of a {cls}.	63.2	66.2	69.7	72.5	75.3
		Template Mining	63.5	67.2	70.3	73.1	75.7
		Hand Engineered [214]	63.7	66.7	70.3	72.9	75.5
	+Flipped View	Hand Engineered [214]	64.1	67.0	70.3	73.0	76.0
Partial	CenterCrop	Classname	62.5	65.7	69.3	72.9	76.2
		a photo of a {cls}.	63.8	66.8	69.8	73.4	76.7
		Template Mining	64.3	67.1	70.3	73.5	76.5
		Hand Engineered [214]	64.6	67.2	70.2	73.7	76.9
	+Flipped View	Hand Engineered [214]	64.7	67.7	70.6	73.8	77.2

Table 2.6: **Augmentation for cross-modal adaptation.** We evaluate the impact of selected augmentation techniques following the same CoOp protocol as in Table 2.1.

This allows us to speed up training by pre-extracting features, resulting in rather fast training speeds.

2.6 Discussion and Limitations

We show that cross-modal training is a lightweight and effective approach for adapting pre-trained multimodal models for downstream uni-modal tasks. One reason for its effectiveness is that it naturally addresses the underspecification of few-shot learning. In the context of vision-language adaptation, one can achieve SOTA results by using existing text labels as free training samples. In the context of vision-audio adaptation, one can learn better visual object classifiers by listening to object sounds (and better audio classifiers by looking at objects!). One attractive aspect of cross-modal learning is that the learned models naturally apply to multimodal test data, such as the classification of videos that contain both visual and audio signals. However, cross-modal learning is less effective when model representations are not well-aligned or insufficiently trained. Nevertheless, due to its simplicity and effectiveness, we hope cross-modal learning becomes a tool for future research on multi-modal adaptation.

2. Multimodality Helps Unimodality: Cross-Modal Few-Shot Learning with Multimodal Models

Method	Source	Target			
	ImageNet	-V2	-Sketch	-A	-R
ResNet50					
Zero-Shot CLIP	58.2	51.3	33.3	21.7	56.0
Linear Probing	55.9	46.0	19.1	12.7	34.9
CoOp (M=4)	63.0	55.1	32.7	22.1	55.0
CoOp (M=16)	63.3	<u>55.4</u>	<u>34.7</u>	23.1	56.6
WiSE-FT ($\alpha=0.5$)	62.9	54.2	33.3	20.3	<u>57.4</u>
Cross-Modal WiSE-FT ($\alpha=0.5$)	65.2	56.6	35.6	<u>22.6</u>	59.5
Cross-Modal Linear Probing	<u>64.5</u>	55.3	33.1	20.0	56.4
ViT-B/16					
Zero-Shot CLIP	66.7	60.8	46.2	47.8	74.0
Linear Probing	65.9	56.3	34.8	35.7	58.4
CoOp (M=4)	71.9	64.2	46.7	48.4	74.3
CoOp (M=16)	71.7	64.6	47.9	49.9	75.1
CoCoOp	71.0	64.1	48.8	50.6	76.2
WiSE-FT ($\alpha=0.5$)	<u>73.0</u>	<u>65.2</u>	<u>49.1</u>	49.8	<u>77.6</u>
Cross-Modal WiSE-FT ($\alpha=0.5$)	72.9	65.4	49.2	<u>50.5</u>	77.8
Cross-Modal Linear Probing	73.2	64.8	47.9	48.3	76.4

Table 2.7: **Robustness under test-time distribution shifts.** We follow CoOp [221]’s protocol for evaluating the test-time performance on variants of ImageNet. We report results with two image encoders (ResNet50 and ViT-B/16), and mark the **best** and second best results. Salient conclusions: (a) Cross-modal linear probing is much more robust than its uni-modal counterpart while being competitive to previous SOTA methods such as WiseFT and CoOp, and (b) it can be further augmented with post-hoc modification through WiseFT to achieve new the SOTA.

Method	Iteration	Time	Accuracy	Gain
Zero-shot CLIP [154]	0	0	60.33	0
Image-Only Linear	12k	15sec	56.44	-3.89
CoOp [221]	100k	14h 40min	62.95	+2.62
ProGrad [221]	100k	17hr	63.45	+3.12
Tip-Adapter [214]	10k	5min	65.18	+5.18
Cross-Modal Linear	12k	15sec	64.51	+4.14
Cross-Modal Partial	12k	2.5min	65.95	+5.57

Table 2.8: **Efficiency and accuracy for different methods on ImageNet-16-shot.** All experiments are tested with batch size 32 on a single NVIDIA GeForce RTX 3090 GPU. Our approaches take less time and achieve SOTA performance.

2. Multimodality Helps Unimodality: Cross-Modal Few-Shot Learning with Multimodal Models

Chapter 3

Language Models as Black-Box Optimizers for Vision-Language Models

Vision-language models (VLMs) pre-trained on web-scale datasets have demonstrated remarkable capabilities on downstream tasks when fine-tuned with minimal data. However, many VLMs rely on proprietary data and are not open-source, which restricts the use of white-box approaches for fine-tuning. As such, we aim to develop a black-box approach to optimize VLMs through **natural language prompts**, thereby avoiding the need to access model parameters, feature embeddings, or even output logits. We propose employing chat-based LLMs to search for the best text prompt for VLMs. Specifically, we adopt an automatic “hill-climbing” procedure that converges to an effective prompt by evaluating the performance of current prompts and asking LLMs to refine them based on textual feedback, all within a conversational process without human-in-the-loop. In a challenging 1-shot image classification setup, our simple approach surpasses the white-box continuous prompting method (CoOp) by an average of 1.5% across 11 datasets including ImageNet. Our approach also outperforms both human-engineered and LLM-generated prompts. We highlight the advantage of conversational feedback that incorporates both positive and negative prompts, suggesting that LLMs can utilize the implicit “gradient” direction in textual feedback for a more efficient search. In addition, we find that the text prompts

generated through our strategy are not only more interpretable but also transfer well across different VLM architectures in a black-box manner. Lastly, we demonstrate our framework on a state-of-the-art black-box VLM (DALL-E 3) for text-to-image optimization.

3.1 Introduction

Vision-language models [3, 105, 153, 186] (VLMs) excel at a wide range of classic vision and multimodal [6, 30, 51, 112, 205] tasks, surpassing the performance of their fully-supervised counterparts on downstream tasks even when fine-tuned with minimal data [114, 219]. However, fine-tuning VLMs typically requires transparent *white-box* access to the model weights, such as gradient-based approaches that rely on backpropagation.

VLMs as black-box services. Despite community efforts to collect web-scale public datasets [165, 166] and to replicate proprietary VLMs [7, 79], an increasing number of models [3, 11, 39, 139, 186, 206] are not releasing their weights due to privacy and legal concerns [108, 124]. Therefore, one cannot use popular *white-box* fine-tuning strategies (such as LoRA [75] and Adapter [73]) that rely on model weights, feature embeddings, and output logits. Given that contemporary black-box VLMs [139, 141] like DALL-E [11, 155] still offer a language-based user interface and may be accessed through APIs that facilitate input and output in *natural language*, this allows users to customize these models through optimizing textual prompts.

Manual prompting. Manual prompt engineering has been proven successful in adapting black-box LLMs to language tasks [92, 193]. Similarly, carefully crafted prompts can enhance the performance of VLMs. For instance, CLIP has demonstrated improved zero-shot recognition performance using specifically tailored prompts, such as "a photo of a {class}" for Internet photos and "a satellite image of a {class}" for satellite imagery. Despite its effectiveness, manual prompting can be a laborious process, inspiring efforts to explore automated prompt creation and thereby remove the need for human involvement. These strategies typically leverage an LLM as a knowledge base to create rich visual descriptors that augment the prompts for each class [129, 149] in a zero-shot fashion.

Human-free prompting with conversational LLMs (our approach). We show how to effectively leverage chat-based LLMs [139] to emulate human-level prompt engineering *without* any human input. We first address an illustrative low-shot image classification task, aiming to find the best class-agnostic prompt (or “template”) for image classification with CLIP. We start with a random set of prompts and evaluate the one-shot training accuracy of each. Then, akin to human prompt engineering, our method repeatedly presents ChatGPT with the best and worst prompts, asking it to review the results and suggest an improvement (see [Figure 3.1](#)).

Learning with implicit “gradients” provided through conversational feedback. One of our key findings is that LLMs can learn the difference between effective and ineffective prompts, and can use this implicit “gradient” direction provided through language to perform more efficient searches. Compared to previous automatic prompting methods that only use LLMs as a knowledge base [129, 149] or paraphrasing tool [222], we show a novel use of LLMs as an *optimizer* that can utilize the patterns hidden in textual feedback. In our experiments, we find that the inclusion of such feedback greatly improves the efficiency and accuracy of our method, sometimes surpassing existing white-box methods [194, 219] on challenging one-shot scenarios.

Optimizing text-to-image generation with DALL-E 3. We further demonstrate our optimization framework on a state-of-the-art black-box VLM, DALL-E [11], for two illustrative one-shot generative tasks: (1) Text-to-image (T2I) generation (see [Figure 3.3](#)), where we sample challenging text queries from Winoground [182] that involve reasoning over compositions of objects, attributes, and relations. Examples include “**an animal watches a person**” and “**there is less milk than orange juice**”, which DALL-E 3 might initially fail to generate. (2) Prompt inversion (see [Figure 3.4](#)), which attempts to reverse-engineer the textual prompt to generate a specific image for later customization [159] (see [Table 3.5](#)). To achieve this, we leverage conversational feedback from a multimodal LLM (GPT4-V [139]) to iteratively refine the prompts based on the current generated images. We present qualitative results in [Table 3.4](#) and conduct a user study to demonstrate that our framework can be more efficient than manual prompting, even for graphical designers experienced with AI content-generation tools.

Our contributions. In this work, we introduce a novel prompting method for VLMs, utilizing an LLM as an *optimizer*. Our black-box approach can surprisingly compete with various white-box methods in a low-shot setting. Additionally, we extensively explore various strategies for conversing with ChatGPT, uncovering several key factors that significantly enhance the efficiency of this tool. We also show that our discovered natural language prompts are not only *interpretable* but also *transfer* better across CLIP architectures, eg., from RN50 to ViT/B-16, than continuous prompts discovered by previous white-box prompting method [219]. Finally, we show practical applications of our framework on text-to-image generation using black-box DALL-E 3. We release our code for future research on prompt optimization and AI-driven content creation ¹.

3.2 Related Works

LLMs for multimodal tasks. Cutting-edge LLMs like GPTs [139, 141] have been successfully applied to multimodal tasks, either through zero-shot composition with pre-trained multimodal models [107, 209] or by jointly fine-tuning with modality-specific encoders [3, 105] on large-scale multimodal datasets [166]. LLMs are also utilized as neuro-symbolic reasoners [53, 120, 170, 218], translating natural language instructions into modular programs (like Python code) that invoke APIs of multimodal models. In this work, we show the potential of LLMs as a *black-box optimizer* for multimodal foundation models with language interfaces, and more specifically vision-language models (VLMs).

Prompt optimization of foundation models. Following the success of in-context learning [17], which appends user-generated natural language instruction and few-shot samples to text inputs, prompting [116] has emerged as the preferred fine-tuning paradigm for LLMs due to its superior performance and parameter-efficiency. However, recent prompt optimization methods, including continuous prefix-tuning [22, 110, 179, 180, 203] and discrete token-searching [31, 35, 171], still operate in a white-box manner, requiring access to either the tokenizer or output logits. Moreover, black-box prompting methods, such as heuristic-based editing [131, 148],

¹Project site: llm-can-optimize-vlm.github.io

are tailored towards language-only tasks and are thus not applicable in VLM settings.

LLMs for prompt optimization. APE [222] leverages an LLM to automatically write prompts using few-shot samples based on instruction induction [72] and paraphrasing [132, 160]. However, it is only designed to address language tasks, while we focus on multimodal tasks using black-box VLMs. LLMs have also proven to be an effective external knowledge base [129, 149, 169] for generating prompts in a zero-shot setting for multimodal models. For example, DCLIP [129] uses GPT3 to come up with rich visual descriptions to improve zero-shot classification with CLIP [153]. We extend this line of work to show that LLMs can *iteratively* optimize prompts for VLMs in a black-box fashion given few-shot samples. We further illustrate that prompt optimization with LLMs can be made more efficient by leveraging *conversational* feedback, such as providing ChatGPT with explicit language feedback on how well the most recent prompt performs. Our findings align with the perspective [28] of LLMs as meta-optimizers that can implicitly perform gradient search through in-context learning.

Few-shot adaptation of VLMs. Prompting has also been successfully adopted in VLMs [44], as demonstrated by methods like CoOp [219] that fine-tune an ensemble of continuous prefix tokens using cross-entropy loss. [114] achieves state-of-the-art few-shot performance with a cross-modal (image and text) cross-entropy loss. However, these methods all require access to model parameters for gradient backpropagation. We also note that while some concurrent works, such as BlackVIP [138] and LFA [140], claim to operate in a “black-box” setting, they still require access to *privileged* information including output logits and embeddings. In this work, we introduce a truly black-box and gradient-free approach that yields competitive results to white-box approaches in extremely low-shot scenarios.

3.3 Prompting VLMs Using Chat-Based LLMs

We now present our approach for prompting VLMs using chat-based LLMs as optimizers.

Preliminaries. Motivated by recent proprietary VLMs [11, 139], we adopt a stricter yet practical black-box setting compared to prior works [138, 140], requiring *minimal* knowledge about the model’s inner workings. This is crucial since releasing

output logits or embeddings can potentially facilitate unauthorized knowledge extraction through distillation methods [69]. Our objective is to enhance the performance of a VLM equipped with a language interface capable of processing a textual prompt $p \in T$. We assume that the targeted task is accompanied by a training dataset denoted as $D_{train} \subset D$, and its performance can be evaluated with respect to the prompt, represented as a function $F : D \times T \rightarrow \mathbb{R}$. For example, in a classification task, $D_{train} = \{x, y\}_n$ where x is an image and y is its class label. The black-box VLM takes the image as input and returns a predicted label. We measure the performance of the textual prompt by calculating the average classification accuracy as $F(D_{train}, p)$. Our goal in prompt engineering is to search for the optimal prompt p^* without accessing or modifying the black-box VLM.

Background: human prompt engineering. Our method draws inspiration from the typical workflow of human prompt engineers. Prompt engineering is often an iterative process that involves: (a) creating an initial prompt $U = \{p_1\}$ based on the understanding of a task, (b) evaluating the performance of prompts in U , (c) refining prompts based on the outcomes, (d) repeating the last two steps until convergence, and (e) returning the prompt p^* with the highest $F(D_{train}, p^*)$. This hands-on approach helps optimize the model’s performance, but it can be tedious and labor-intensive. Algorithm 1 formally illustrates this process.

Example: prompting for image classification with CLIP [153]. CLIP is one of the most popular VLM that takes a set of class-specific prompts when performing “zero-shot” image classification. [153] details the laborious prompting procedure over the course of a year. Interestingly, they find that a default class-agnostic prompt (or so-called “template”), “a photo of a {class}” can provide a decent boost in accuracy for most datasets compared to using vanilla class labels. In this scenario, the evaluation function F is the classification accuracy on the test set, and the prompt $p = \{\text{“a photo of a {class}”} \mid c \in C\}$, where C is the set of class names for a given dataset.

Prompting with chat-based LLMs (our approach). Given the strong in-context reasoning capabilities of LLMs, we envision them as a *black-box optimizers* that can improve prompts based on their performance outcomes, akin to how human prompt engineers iteratively refine prompts. Specifically, we maintain a pool of prompts U and their corresponding performance outcomes S . In each iteration, we

Algorithm 1 We formalize *human* prompt engineering with the following algorithm, which motivates our LLM-based algorithm (2).

Require: $D_{\text{train}} = \{x, y\}_n$: training samples, $F : D \times T \rightarrow \mathbb{R}$: evaluation function

- 1: Create initial prompts: $\mathcal{U} \leftarrow \{p_1\}$
- 2: Evaluate prompts on training set: $S \leftarrow \{F(D_{\text{train}}, p_1)\}$
- 3: **while** not converged **do**
- 4: Generate a new prompt p' based on S
- 5: Evaluate the new prompt: $s' = F(D_{\text{train}}, p')$
- 6: $\mathcal{U} \leftarrow \mathcal{U} \cup \{p'\}$
- 7: $S \leftarrow S \cup \{s'\}$
- 8: **end while**
- 9: **return** optimal prompt $p^* \leftarrow \arg \max_{p \in \mathcal{U}} F(D_{\text{train}}, p)$

provide the LLM with both *positive* and *negative* prompts, such as the highest and lowest-performing candidates. Such textual feedback through in-context prompts offers LLMs an implied "gradient" direction [28], making optimization more efficient than taking random local steps. We facilitate this feedback mechanism through *conversations* with state-of-the-art chat-based LLMs like ChatGPT [141] as illustrated in Figure 3.1.

3.4 Illustrative Few-Shot Classification Task

We illustrate our approach using a few-shot image classification task. Specifically, a prompt $p \in T$ consists of a set of class-specific prompts – that is, one textual description per class. The evaluation function F takes the prompt p , along with an image dataset D_{train} , and returns the accuracy using the black-box VLM. To prevent overfitting and simplify our search space, we restrict our search to finding a single class-agnostic template, e.g., a **photo of a {}**, filling in the blank with label names provided with the dataset.

Outline of our approach (Alg. 2). To start, we sample entirely random initial prompts from a text corpus such as LAION-COCO [165] captions. Our approach follows the classical *stochastic hill-climbing framework with random-restart* [160], which prevents ChatGPT from being trapped in local optima by balancing "exploration" and "exploitation". Our **restart** mechanism is implemented by sampling n_{restart}

Algorithm 2 LLM-based prompt engineering on the illustrative classification task. Our algorithm requires a chat-based LLM and a (black-box) evaluation function, such as accuracy. We highlight mechanisms for “exploration” (**restart** and **reset**) in blue and “exploitation” (**iter**) in red. We mark the key component of “**conversational feedback**” of our approach in violet.

Require: $D_{\text{train}} = \{x, y\}_n$: training samples, $F : D \times T \rightarrow \mathbb{R}$: evaluation function.
Require: n_{restart} : number of initial sampled prompt sets, n_{reset} : number of resets for a prompt set, n_{iter} : number of hill-climbing iterations, m : size of one initial prompt set, k : number of prompts send to ChatGPT.

- 1: $p^* \leftarrow \emptyset$
- 2: **for** $1::n_{\text{restart}}$ **do**
- 3: Sample a new prompt set, $\mathcal{U}_{\text{init}} \leftarrow \{p_1, \dots, p_m\}$
- 4: **for** $1::n_{\text{reset}}$ **do**
- 5: Reset to initial prompt set: $\mathcal{U} \leftarrow \mathcal{U}_{\text{init}}$
- 6: **for** $1::n_{\text{iter}}$ **do**
- 7: Sort \mathcal{U} by score outcomes $\{F(D_{\text{train}}, p)\}_{p \in \mathcal{U}}$
- 8: $\mathcal{U}_{\text{top}} \leftarrow$ top-k prompts in \mathcal{U}
- 9: $\mathcal{U}_{\text{bot}} \leftarrow$ bottom-k prompts in \mathcal{U}
- 10: **Get a new prompt** $p_{\text{new}} \leftarrow \text{LLM}(\mathcal{U}_{\text{top}}, \mathcal{U}_{\text{bot}})$
- 11: $\mathcal{U} \leftarrow \mathcal{U} \cup \{p_{\text{new}}\}$
- 12: **end for**
- 13: $p^* \leftarrow \arg \max_{p \in \mathcal{U} \cup \{p^*\}} F(D_{\text{train}}, p)$
- 14: **end for**
- 15: **end for**
- 16: **return** prompt with highest score p^*

initial prompt sets to encourage exploration. Because ChatGPT performs stochastic top-k sampling for text generation (as we adopt the default temperature of 1.0), we also implement a **reset** mechanism to foster additional exploration by retrying a given prompt set n_{reset} times. For exploitation, we converse with ChatGPT for n_{iter} iterations. Lastly, we present ChatGPT both the top and bottom-performing prompts, denoted as $(\mathcal{U}_{\text{top}}, \mathcal{U}_{\text{bot}})$. We show that this simple adjustment can improve the efficiency of our approach in [Figure 3.2](#).

Experimental setup. We apply our approach to the few-shot image classification benchmark introduced in CoOp [219], which is the most commonly studied setup for fine-tuning VLMs. This benchmark involves a collection of 11 datasets covering diverse image domains including ImageNet [30] and more niche datasets such as

FGVC-Aircraft [125]. For each dataset, we adhere to the same three-fold k-shot train sets in [114], reporting the average accuracy across all folds. Importantly, our method only utilizes the train set to compute the score and does not require the few-shot validation set. We use CLIP following prior work [114, 219] to emulate a black-box VLM, and we employ ChatGPT (GPT3.5) as the chat-based LLM.

Implementation details. To start, we sample entirely random 1M initial prompts from a text corpus (LAION-COCO [165] captions). For each caption, we extract all the noun phrases using spaCy part-of-speech tagging [71]. Subsequently, we replace one noun phrase in the caption with ‘‘{}’’ (a placeholder where the class name will be inserted) to create a template. Given that each caption contains an average of 2 noun phrases, our initial prompt pool consists of approximately 2M templates. We run our algorithm with $n_{\text{restart}} = 20$ restarts, $n_{\text{restart}} = 50$ resets, and $n_{\text{restart}} = 10$ iterations. We opt to sample $m = 100$ prompts per restart and present the top and bottom $k = 15$ prompts to ChatGPT. We adopt `gpt-3.5-turbo-0301` model for ChatGPT using OpenAI’s official API and keep the default sampling temperature of 1.0. For a fair comparison, we use CLIP-RN50 for our experiments following prior work [114, 219]. We will open-source our code and release the initial prompt pool (LAIONCOCO-1M) to the public.

Oracle white-box baselines. Our black-box setup substantially differs from, and is more constrained than, the scenarios considered in previous white-box baselines. Specifically, we do **not** expose the pre-trained weights, model architectures, feature embeddings, or even output logits of VLMs. These constraints render many established *gradient-based fine-tuning* baselines inapplicable. Among the *oracle* white-box approaches we later compare to, **CoOp** [219] performs continuous prompting and requires backpropagation across all layers. **WiSE-FT** [194] ensembles fine-tuned weights with the original CLIP weights. **Cross-Modal Adaptation** [114] fine-tunes a linear classifier leveraging both image and text embeddings from CLIP. Finally, while **DCLIP** [129] queries GPT3 for rich visual descriptors for each class and does not require gradient-based fine-tuning, it performs *prompt ensembling* using 4-6 class-specific prompts, which breaches our black-box assumption for accessing the output logits.

Black-box methods. We additionally benchmark our method against truly black-box solutions, including the vanilla class-agnostic templates “{class}” and “a photo

of a `{class}`”. Also, we compare our approach to the best **Hand-Engineered** templates released by OpenAI, searched using *test set* performance to represent the theoretical upper bound of human performance, eg., “a **centered satellite photo of {class}**.” for EuroSAT [61]. Finally, we present two versions of conversational feedback of our approach: (a) using 30 positive (**P only**) or (b) using 15 positive and 15 negative prompts (**P+N**) in each iteration. For a fair comparison, both of our approaches start with the same initial sampled prompts, referred to as **LAIONCOCO-1M**. We also show the performance of the best initial sampled prompt searched using trainset performance.

SOTA one-shot performance against existing methods on 11 datasets.

We report the test set performance of our method versus the aforementioned baselines in a challenging 1-shot classification scenario in [Table 3.1](#). First, compared to the top-performing initial prompts selected from **LAIONCOCO-1M** based on train set performance, our prompt optimization using ChatGPT notably improves the initial prompts by an average of 5% (56% to 61%). Remarkably, our black-box approach surpasses the two white-box gradient-based fine-tuning techniques CoOp and WiSE-FT by at least 1.5%. Given that both CoOp and our method optimize a single class-agnostic template, we attribute this gap in performance to *reduced overfitting*. More specifically, we posit that our optimization space of natural language effectively acts as a regularizer in extremely low-shot tasks, standing as a more robust alternative to the continuous prompting approach of CoOp. Furthermore, our method benefits from textual feedback and shows improved performance by 1.0% when using both positive and negative prompts.

Incorporating negative prompts leads to more efficient optimization.

In [Figure 3.2](#), we demonstrate that incorporating both positive and negative prompts fosters better optimization efficiency, achieving higher accuracy within a much fewer number of resets. Specifically, we hypothesize that LLMs can leverage the implicit “gradient” direction suggested in textual feedback to achieve faster convergence.

3.5 More Benefits of Natural Language Prompts

In this section, we delve deeper into the advantages of utilizing natural language prompts compared to the continuous prompts [219]. We highlight that the prompts

derived through our method are *interpretable*; for instance, they often contain descriptions of the targeted image domain. Our prompts can also *transfer* across CLIP architectures in a *black-box manner*, such as from RN50 to ViT/B-16.

Interpretable natural language prompts. While CoOp [219] concedes that continuous prompts can be difficult to interpret, our method – without explicitly instructing ChatGPT to generate interpretation – often yields interpretable results. Table 3.2 showcases the templates returned by our algorithm for each dataset, frequently including keywords that reflect the targeted image domain. For example, the template for Food101 [14] mentions “diverse cuisine and ingredients”, and the template for UCF101 [178] (an action recognition dataset) mentions “in motion”. Likewise, these templates identify general stylistic attributes of the datasets; they refer to “bright and natural lighting” for ImageNet [30] and note images that “emphasize the subject” for Caltech101 [98]. These prompts are particularly intriguing because we do not provide ChatGPT with any information about the downstream task, yet it manages to generate prompts containing domain-specific keywords that are similar to those engineered by human experts.

Black-box prompt transfer. Our text prompts also maintain consistently high performance across different CLIP backbones. For comparison, since CoOp uses the same tokenizer for all CLIP architectures (including RN50, RN101, ViT/B-32, and ViT/B-16) and optimizes continuous prompts of the same shape (16 x 512), we assess the transferability of these learned continuous prompts from RN50 to other backbones using the official weights on 16-shot ImageNet. Table 3.3 showcases the results of this experiment, where we also include the baseline prompt “a photo of a {}” for reference. We observe a significant decline in accuracy when transferring CoOp’s prompts (up to a 40% decrease despite utilizing more powerful backbones), implying that continuous prompts tend to overfit to the specific CLIP model. In contrast, our natural language prompts maintain their performance and outperform the baseline across all backbones.

3.6 Application: Text-to-Image Generation

In this section, we present a direct application of our prompt optimization framework to generative tasks using a truly black-box text-to-image (T2I) VLM, DALL-E 3 [11].

Optimizing T2I using a multimodal LLM. DALL-E 3 can generate high-fidelity images following diverse user queries, but crafting effective prompts is tricky even for designers experienced with AI content generation tools [117]. Therefore, we are motivated to implement our LLM-based optimization framework to assist with creative visual design. Our framework is shown in [Figure 3.3](#) for the illustrative task of text-to-image generation. In this task, the user specifies a query (topic) in text, such as “an animal watches a person”, and the goal is to write a prompt that can generate an image reflecting this topic. We adopt a *multimodal* LLM GPT4-V [139] (`gpt-4-1106-preview`) to provide feedback on the generated image and optimize the prompt. We find that this framework is surprisingly effective due to GPT4-V’s strong visual reasoning capabilities, which can often spot subtle errors in generated images and offer more accurate prompts.

Task setup. For *T2I generation*, we experiment with a subset of 100 text queries from Winoground [182] that involve complex attribute and relation reasoning, which DALL-E might initially fail to generate. Our framework refines the prompts to capture the user-specified topics using a few (three) iterations. We also attempt a reverse task of *prompt inversion*: given a user-specified reference (query) image, our framework reverse-engineers the prompt to have DALL-E generate the same object or scene in the query image (see [Figure 3.4](#)). This enables users to easily make customizations [159] (see [Table 3.5](#)), such as having the character in a reference image perform various actions or change scenes. For this task, we sample 100 random queries from DiffusionDB [192]. We provide qualitative results in [Table 3.4](#). We hire two volunteers to assess the faithfulness of the images generated by our method, and to compare these with the images manually prompted by two designers (each with one year of experience in AI content generation), as shown in [Table 3.6](#).

Remarks on limitations. While we show promising results, we also note some failure cases due to the inherent limitations of foundation models. For example, GPT4-V might fail to describe abstract and artistic details, and DALL-E 3 often fails to generate the correct number of objects. We believe that our framework can benefit from more capable foundation models in the future.

3.7 Discussion and Limitations

Summary. We present the first attempt to leverage LLMs as prompt engineers for VLMs. On the well-studied setup of one-shot image classification, our method surpasses existing human-engineered prompts and even rivals white-box approaches. Central to the success of our method is the utilization of conversational feedback, enabling chat-based LLMs to efficiently steer VLMs in the right direction. This process leads to a set of interpretable prompts bearing considerable resemblance to those crafted by humans. Importantly, our natural language prompting setup is a lot more constrained than the assumed scenarios of previous white-box or even some black-box settings [138], because we do not expose the model weights and outputs of VLMs. We finally apply our framework to illustrative generative tasks using a truly black-box text-to-image VLM (DALL-E 3).

Limitations and future work. As with any work utilizing LLMs, there are various ethical concerns, including biases in the LLM’s output prompts. Moreover, while we try to minimize the overall cost and the total number of API calls, the energy consumption associated with LLMs remains a substantial concern. It is vital to note that we do not intend to compete directly with white-box baselines that can improve visual and text representations with more data. Lastly, we are limited to costly human evaluation for T2I generation in this study. Future work may adopt automatic evaluation [24, 68, 76, 113] for large-scale experiments.

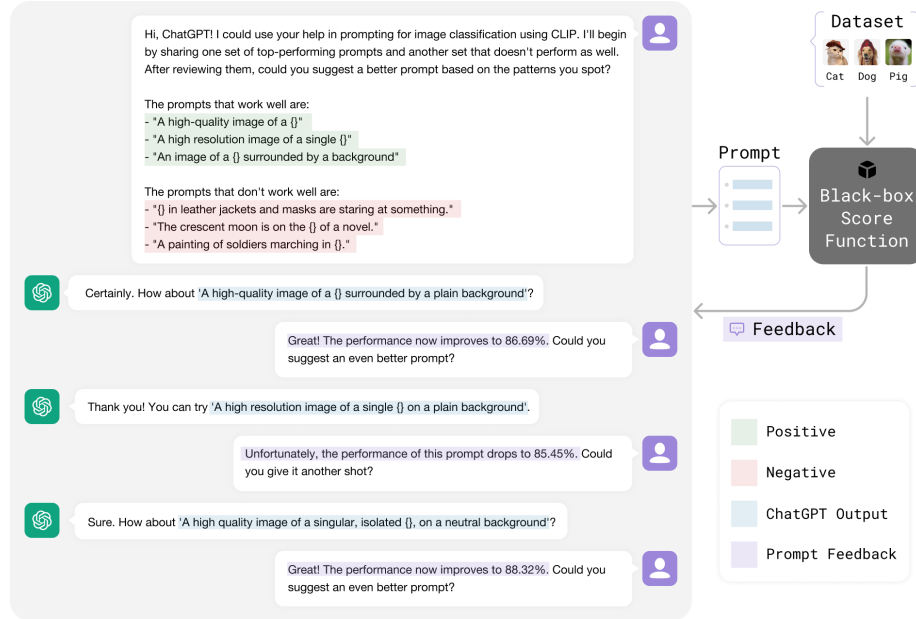


Figure 3.1: **Prompting VLMs using chat-based LLMs.** Similar to how human prompt engineers iteratively test and refine prompts, we employ ChatGPT [139, 141] to continuously optimize prompts for vision-language models (VLMs). Our iterative approach assesses the performance of ChatGPT-generated prompts on a few-shot dataset (highlighted in blue) and provides feedback (marked in violet) to ChatGPT through simple conversations, as depicted in the illustrative figure. This straightforward method delivers state-of-the-art results for one-shot image classification across 11 datasets using CLIP, operated in a black-box manner without accessing model weights, feature embeddings, or output logits. We show that providing both positive (in green) and negative prompts (in red) enhances efficiency. Remarkably, our approach outperforms both white-box methods such as gradient-based continuous prompting (CoOp [219]) and human-engineered prompts [153] in this extremely low-shot scenario. This figure only shows a typical conversation using ChatGPT’s web user interface. Our code implementation follows this pattern using the ChatGPT API.

3. Language Models as Black-Box Optimizers for Vision-Language Models

Method	Dataset											Avg
	Caltech	ImageNet	Aircraft	Food	Pets	Cars	SUN	UCF	DTD	EuroSAT	Flowers	
Oracle white-box approaches												
Cross-Modal [114]	89.1	61.6	20.6	77.1	85.7	59.0	63.4	64.7	49.9	61.8	76.3	64.7
WiSE-FT [194]	85.5	58.3	18.6	71.9	81.7	55.7	56.6	59.4	44.2	52.3	65.8	59.1
CoOp [219]	87.5	57.2	9.6	74.3	85.9	55.6	60.3	61.9	44.4	50.6	68.1	59.6
DCLIP [129]	-	59.6	-	76.4	83.8	-	-	-	41.7	34.7	-	-
Manual prompting approaches												
{}	78.5	55.3	15.5	74.0	78.9	52.2	53.4	55.5	41.4	32.1	57.3	54.0
a photo of a {}	84.5	57.9	15.9	74.0	83.2	53.9	58.0	56.9	38.8	28.6	60.2	55.6
Hand-Engineered [153]	86.3	58.2	17.3	77.3	<u>85.8</u>	55.6	58.5	61.5	42.3	37.6	66.1	58.8
Our black-box approaches												
LAIONCOCO-1M	81.4	56.2	17.4	76.5	79.6	51.3	54.9	55.8	43.1	38.6	61.3	56.0
Ours (P only)	<u>89.0</u>	<u>59.4</u>	<u>17.9</u>	<u>77.8</u>	85.7	<u>55.7</u>	<u>60.4</u>	58.7	<u>43.6</u>	<u>46.7</u>	<u>66.6</u>	<u>60.1</u>
Ours (P+N)	89.1	59.6	18.1	78.3	88.1	56.2	61.0	<u>60.2</u>	44.8	49.0	67.2	61.1

Table 3.1: **Comparison of our method with other baselines on one-shot classification tasks.** We report the average accuracy of each method across three folds, optimized using 1-shot training sets. We **bold** the best black-box result for each dataset, and underline the second best result. First, we note that our approach can effectively improve upon the initial prompts selected from LAIONCOCO-1M from 56% to 61%. Our approach is also competitive against the best Human-Engineered prompts released by OpenAI [153] searched using *test set* performance. Additionally, we show that using both positive and negative prompts improves the overall accuracy by 1%. For reference, we report *oracle* white-box approaches in gray. Remarkably, we also surpass white-box solutions such as WiSE-FT [194] and CoOp [219] by 1.5%. These methods require either gradient-based fine-tuning (CoOp/WiSE-FT/Cross-Modal) or prompt ensembling using output logits (DCLIP). While our approach is less effective than the SOTA white-box method (Cross-Modal Adaptation), we stress that our black-box setup is significantly more challenging, because we restrict the optimization space to *natural language* and do *not* access the pre-trained weights, model architectures, feature embeddings, and output logits of VLMs.

3. Language Models as Black-Box Optimizers for Vision-Language Models

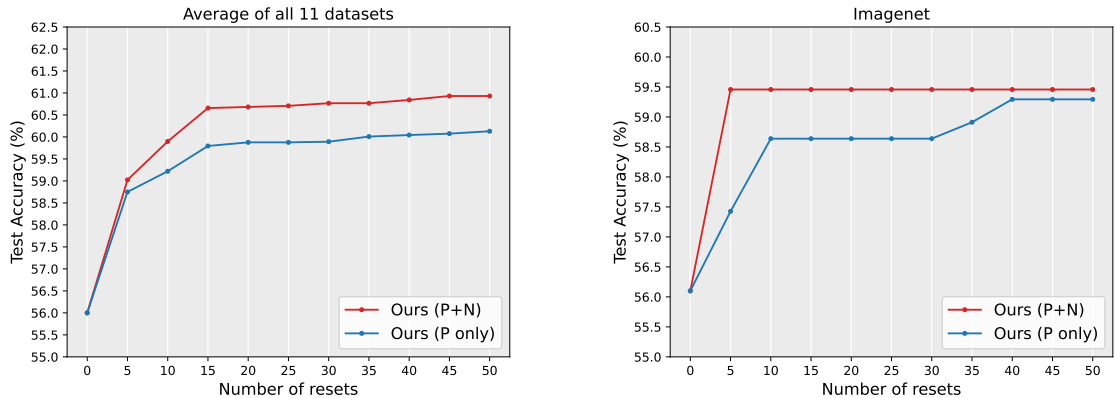


Figure 3.2: **Conversational feedback incorporating both positive and negative prompts leads to improved efficiency.** We fix the number of restarts to 20 and iterations to 10, and ablate different numbers of resets on all 11 datasets (left) and ImageNet (right). Notably, our approach using “P+N” (both top-15 and bottom-15 prompts) can optimize faster within a much fewer number of resets than using “P-Only” (top-30 prompts), resulting in the highest overall performance.

Dataset	Example of Top Templates
Caltech [98]	An image of a {} with a blurred background that emphasizes the subject
DTD [26]	The essential elements of {} are amplified with visual simplicity
EuroSAT [61]	A top-down view of {} arranged in a pattern {}
Aircraft [125]	A clear, high-quality image of a single {} with a white background
Food [14]	A {} featuring diverse cuisine and ingredients
ImageNet [30]	An image of a {} with bright and natural lighting
Flowers [136]	A clear and vivid photograph of the {} in its natural setting
Pets [145]	A {} with distinct and recognizable characteristics
Cars [94]	A {} featuring a wide range of color options for easy selection
SUN [200]	A high-resolution photo of a {} with clear background and natural lighting
UCF [178]	A black and white photo of a {} in motion

Table 3.2: **Example templates returned by our algorithm on each dataset.** Although we do not provide ChatGPT with any information regarding the targeted dataset, we observe that the resulting templates are remarkably similar to human-engineered templates, with many domain-specific details such as “motion” and “cuisine”, and stylistic elements such as “bright and natural lighting”.

Method	RN50	→RN101	→ViT-B/32	→ViT-B/16
a photo of a {}	57.9	60.6	61.9	66.6
CoOp	63.0	20.6	31.7	39.5
Ours	59.9	60.7	62.2	67.0

Table 3.3: **Black-box prompt transfer from ResNet-50 to other CLIP architectures.** We evaluate both our natural language prompts and CoOp’s continuous prompts on 16-shot ImageNet, which are trained using the RN50 CLIP backbone. As a reference point, we include the baseline prompt “a photo of a {}”, and show that the prompts derived from our method using RN50 consistently surpass it after transferring to different backbones. In contrast, while CoOp achieves better 16-shot ImageNet performance using RN50, its performance plummets during the transfer, e.g., from 63% to a mere 21% for RN101.

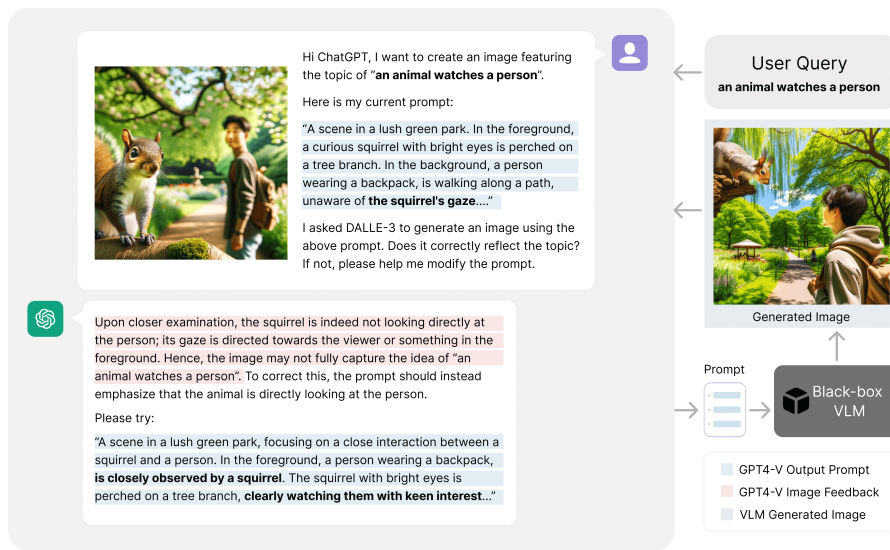


Figure 3.3: **Improving text-to-image (T2I) generation using chat-based multimodal LLMs.** We apply our framework to optimize prompts for the state-of-the-art black-box generative VLM, DALL-E 3 [11], using the multimodal GPT4-V [139]. For complicated user queries that DALL-E 3 may initially fail to generate, we send the generated image (in violet) along with the current prompt to GPT4-V to ask for feedback on improvements (in red) and then generate a new prompt (in blue). We show that such a simple framework is surprisingly effective at correcting DALL-E 3 mistakes on some challenging Winoground [182] text queries that involve action, logical, and spatial reasoning. We conduct a human evaluation on the quality of generated images in Table 3.6. We open-source our code at link to facilitate future research on AI-driven content generation.

3. Language Models as Black-Box Optimizers for Vision-Language Models

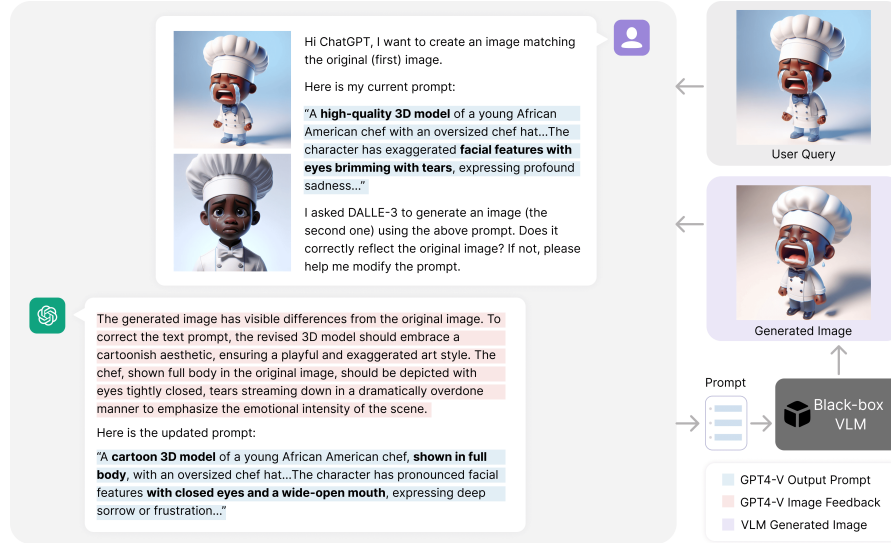


Figure 3.4: **Prompt inversion using chat-based multimodal LLMs.** We apply our framework to reverse engineer the text prompt to generate the same user-queried image. We send the generated image (in violet) along with the original image to GPT4-V to ask for feedback on improvements (in red) and then generate a new prompt (in blue).

User Query	Init. Image	LLM Feedback	Final Image
Text-to-image generation			
There is less milk than orange juice.		Incorrect, the milk bottle appears full, more than orange juice...	
A shorter person is covering the eyes of a taller person.		Incorrect, the taller person is covering the shorter person's eyes. Instead, ...	
Prompt inversion			
		The scarf should feature red and white stripes, and the fur is fluffy...	
		The coat should be buttoned and the lighting exhibits a stronger contrast...	

Table 3.4: **Examples of T2I optimization.** We show that our framework (Figure 3.3) can automatically improve the faithfulness of images generated by DALL-E 3, with respect to user-specified textual topics (for T2I generation) or reference images (for prompt inversion). This is achieved through three rounds of prompt optimization, using feedback from the multimodal LLM (GPT4-V).



User Query	Inverted Image	Example 1	Example 2	Example 3	Example 4	Example 5
						
		Give the dog a cat friend.	Make the dog be in the middle of a jump.	Make the dog do a handstand.	Make the dog lie down on its side.	Make the dog swim in water.
						
		Make the owl fight a hawk.	Make the owl flap its wings.	Make the owl fully green.	Make the owl stand in front of the moon.	Make the owl walk in the city.

Table 3.5: **Customization via prompt inversion.** Users can simply append extra descriptions to the inverted prompts to customize their main characters in queried images.

Task	Method	Init. (std)	Final (std)	Δ
Text-to-Image	Human	2.28 (.45)	2.86 (.61)	0.58
	Ours	2.62 (.36)	3.56 (.54)	0.94
Prompt Inversion	Human	1.58 (.48)	2.76 (.53)	1.18
	Ours	1.94 (.39)	3.68 (.47)	1.74

Table 3.6: **Our method enhances faithfulness in T2I generation.** We hire two human annotators to assess the faithfulness of images generated from user queries, e.g., textual topics for Text-to-Image, or reference images for Prompt Inversion. The scores are measured on a 1-to-5 Likert scale, with 1 signifying contradiction and 5 indicating perfect alignment with the user’s goal. Our approach benefits from three iterations of prompt optimization and consistently outperforms human-engineered prompts by designers who have one year of experience in AI content generation.

3. Language Models as Black-Box Optimizers for Vision-Language Models

Chapter 4

Revisiting the Role of Language Priors in Vision-Language Models

Vision-language models (VLMs) are impactful in part because they can be applied to a variety of visual understanding tasks in a zero-shot fashion, without any fine-tuning. We study generative VLMs that are trained for next-word generation given an image. We explore their zero-shot performance on the illustrative task of image-text retrieval across 9 popular vision-language benchmarks. Our first observation is that they can be repurposed for discriminative tasks (such as image-text retrieval) by simply computing the match score of generating a particular text string given an image. We call this probabilistic score the **Visual Generative Pre-Training Score** (VisualGPTScore). While the VisualGPTScore produces near-perfect accuracy on some retrieval benchmarks, it yields poor accuracy on others. We analyze this behavior through a probabilistic lens, pointing out that some benchmarks inadvertently capture unnatural language distributions by creating adversarial but unlikely text captions. In fact, we demonstrate that even a “blind” language model that ignores any image evidence can sometimes outperform all prior art, reminiscent of similar challenges faced by the visual-question answering (VQA) community many years ago. We derive a probabilistic post-processing scheme that controls for the amount of linguistic bias in generative VLMs at test time without having to retrain or fine-tune the model. We show that the VisualGPTScore, when appropriately debiased, is a strong zero-shot baseline for vision-language understanding, oftentimes producing state-of-the-art

accuracy.

4.1 Introduction

Vision-language models (VLMs) trained on web-scale datasets will likely serve as the foundation for next-generation visual understanding systems. One reason for their widespread adoption is their ability to be used in an “off-the-shelf” (OTS) or zero-shot manner without fine-tuning for specific target applications. In this study, we explore their OTS use on the task of image-text retrieval (e.g., given an image, predict the correct caption out of K options) across a suite of 9 popular benchmarks [30, 74, 112, 123, 182, 187, 205, 208, 216].

Challenges. While the performance of foundational VLMs is impressive, many open challenges remain. Recent analyses [88, 208] point out that leading VLMs such as CLIP [153] may often degrade to “bag-of-words” that confuse captions such as “the horse is eating the grass” and “the grass is eating the horse”. This makes it difficult to use VLMs to capture *compositions* of objects, attributes, and their relations. But somewhat interestingly, large-scale language models (LLMs) trained for autoregressive next-token prediction [17] seem to be able to discern such distinctions, which we investigate below. A related but under-appreciated difficulty is that of *benchmarking* the performance of visio-linguistic reasoning. Perhaps the most well-known example in the community is that of the influential VQA benchmarks [6], which could be largely solved by exploiting linguistic biases in the dataset – concretely, questions about images could often be answered by “blind” language-only models that did not look at the image [51]. Notably, we find that such blind algorithms still excel on many contemporary image-text retrieval benchmarks where VLMs may struggle.

Generative models for discriminative tasks. We tackle the above challenges by revisiting the role of language priors through a probabilistic lens. To allow for a probabilistic treatment, we focus on generative VLMs that take an image as input and stochastically generate text via next-token prediction [103, 105]. We first demonstrate that such models can be easily repurposed for discriminative tasks (such as retrieval) by setting the match score for an image-text pair to be the probability that the VLM would generate that text from the given image, or $P(\text{text}|\text{image})$. We call

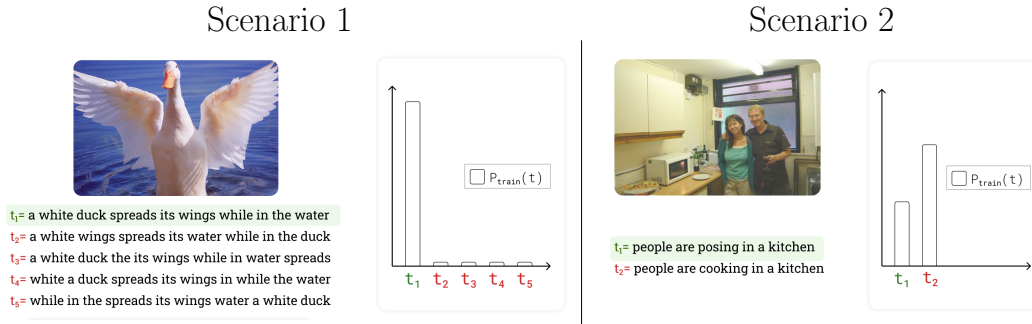


Figure 4.1: **Two train-test shifts encountered in image-to-text retrieval tasks.** Scenario 1 (**left**) constructs negative captions by shuffling words in the true caption (as in ARO-Flickr), but this produces implausible text such as “white a duck spreads its wings in while the water”. Here, exploiting the language bias of the training set will help since it will downweight the match score for such implausible negative captions. In fact, we show that a blind language-only model can easily identify the correct caption. Scenario 2 (**right**) constructs negative captions that are curated to be plausible (as in SugarCreme). Here, the language bias of the training set may hurt, since it will prefer to match common captions that score well under the language prior; i.e., the incorrect caption of “people are cooking in a kitchen” is more likely than the true caption of “people are posing in a kitchen” under the language prior, and so removing the language bias improves performance.

this probability score the Visual Generative Pre-Training Score, or VisualGPTScore. Computing the VisualGPTScore is even more efficient than next-token generation since given an image, all tokens from a candidate text string can be evaluated in parallel. Though conceptually straightforward, such an approach (to our knowledge) has not been proposed in the literature. In fact, the generative VLMs [103] that we analyze train *separate* discriminative heads for matching/classifying image-text pairs, but we find that their language generation head itself produces better scores for matching (since it appears to better capture compositions). Indeed, the OTS VisualGPTScore performs surprisingly well on many benchmarks, even producing near-perfect accuracy on ARO [208]. But it still struggles on other benchmarks such as Winoground [182]. We analyze this below.

The role of language priors. We analyze the discrepancy in performance across benchmarks from a probabilistic perspective. Our key insight is that many benchmark biases can be formalized as mismatching distributions over text between train and test data - $P_{train}(\text{text})$ versus $P_{test}(\text{text})$. We use a first-principles analysis

to account for distribution shift by simply reweighting the VisualGPTScore with the Bayes factor $P_{test}(\text{text})/P_{train}(\text{text})$, a process we call *debiasing*. To compute the Bayes reweighting factor, we need access to both the train and test language prior. We compute $P_{train}(\text{text})$ from an OTS VLM by drawing Monte-Carlo samples of $P_{train}(\text{text}|\text{image})$ from trainset or Gaussian noise images. Because $P_{test}(\text{text})$ may require access to the test set, we explore simplifying assumptions that it is (a) identical to $P_{train}(\text{text})$, (b) uninformative/uniform, or (c) tunable from a held-out val set. Our analysis helps explain the strong performance of the VisualGPTScore on certain benchmarks and its poor performance on others. Moreover, our analysis offers simple strategies to improve performance through debiasing. We conclude by showing a theoretical connection between debiasing and mutual information, which can be seen as a method for removing the effect of marginal priors when computing joint probability scores.

Empirical Analysis. We conduct a thorough empirical evaluation of the OTS VisualGPTScore (and its debiased variants) for open-sourced image-conditioned language models [103, 105] across 9 popular vision-language benchmarks. We first point out that the VisualGPTScore by itself produces SOTA accuracy on certain benchmarks like ARO [208] where their inherent language biases help remove incorrect captions that are also unnatural (such as "a white duck the its wings while in water" as shown in Fig. 4.1). In fact, we show that blind baselines also do quite well on these benchmarks, since language-only models can easily identify such implausible captions. However, such language biases do not work well on benchmarks where incorrect captions are also realistic. Here, VisualGPTScore should be debiased so as not to naively prefer more common captions that score well under its language prior. When given access to a validation set that reveals the amount of language bias in the benchmark, debiasing consistently improves performance on benchmarks such as Flickr30K [205] and Winoground [182]. Interestingly, we find that debiasing can also improve accuracy on the *train* set used to learn the generative VLMs, indicating that such models learn biased estimates of the true conditional distribution $P_{train}(\text{text}|\text{image})$. We describe this further in our appendix.

4.2 Related Works

Vision-language modelling. State-of-the-art VLMs like CLIP [153] are pre-trained on web-scale image-text datasets [165, 166] using discriminative objectives including image-text contrastive (ITC) [82, 153] and image-text matching (ITM) [102, 103] loss, typically formulated as $P(\text{match}|\text{image}, \text{text})$. These pre-trained models exhibit robust zero-shot and few-shot [114, 195] performance on traditional discriminative tasks [30, 112], often on par with fully-supervised models. More recently, image-conditioned language models like Flamingo [3] and BLIP [103, 105] incorporate generative objectives [9] primarily for downstream tasks such as captioning [2] and VQA [51].

Visio-linguistic compositionality. Benchmarks like ARO [208], Crepe [123], Winoground [182], EqBen [187], VL-CheckList [216], and SugarCrepe [74] show that discriminative scores of VLMs, such as ITCScore and ITMScore, fail on their image-text retrieval tasks that assess compositional reasoning. Concurrently, advances on these tasks often involve fine-tuning discriminative VLMs with more data. One of the most popular approaches, NegCLIP [208], augments CLIP using programmatically generated negatives from original texts. Extending this, subsequent studies propose more expensive and heavily-engineered solutions. SyViC [21] fine-tunes VLMs on million-scale synthetic images to augment spatial, attributive, and relation understanding. SGVL [66] and Structure-CLIP [78] sample negatives using costly scene graph annotations. MosaiCLIP [173] and SVLC [37] use linguistic tools such as scene graph parsers and LLMs to design better negative captions. The most recent DAC [38] leverages a combination of foundation models including BLIP2, ChatGPT, and SAM to rewrite and augment image captions.

Generative pre-training and scoring. Vision models trained with *discriminative* objectives often lack incentives to learn structure information [15, 181]. Similarly, early LLMs trained with *discriminative* approaches, such as BERT [32] and RoBERTa [119], have also been criticized as bag-of-words models insensitive to word order [10, 67, 144, 174]. Conversely, generative pre-trained LLMs [152] demonstrate exceptional compositional understanding while pre-trained solely with a next-token prediction [9] loss. Furthermore, generative scores of LLMs [25, 139, 215] have flexible usage in downstream tasks, such as text evaluation [43, 207] and reranking [90].

4.3 The role of language priors

In this section, we present a simple probabilistic treatment for analyzing the role of language priors in image-conditioned language models (or generative VLMs). Motivated by their strong but inconsistent performance across a variety of image-text retrieval benchmarks, we analyze their behavior when there exists a mismatch between training and test distributions, deriving simple schemes for addressing the mismatch with reweighting. We conclude by exposing a connection to related work on mutual information.

Computing $P(\mathbf{t}|\mathbf{i})$. To begin our probabilistic treatment, we first show that image-conditioned language models (that probabilistically generate text based on an image) can be repurposed for computing a score between a given image \mathbf{i} and text caption \mathbf{t} . The likelihood of a text sequence $\mathbf{t} = \{t_1, t_2, \dots, t_m\}$ conditioned on image \mathbf{i} is naturally factorized as an autoregressive product [9]:

$$P(\mathbf{t}|\mathbf{i}) = \prod_{k=1}^m P(t_k | t_{<k}, \mathbf{i}) \quad (4.1)$$

Image-conditioned language models return back m softmax distributions corresponding to the m terms in the above expression. Text generation requires *sequential* token-by-token prediction, since token t_k must be generated before it can be used as an input to generate the softmax distribution over token t_{k+1} . Interestingly, given an image \mathbf{i} and a text sequence \mathbf{t} , the above probability can be computed in *parallel* because the entire sequence of tokens $\{t_k\}$ is already available as input. We provide a visual illustration in [Figure 4.2-a](#).

Train-test shifts. Given the image-conditioned model of $P(\mathbf{t}|\mathbf{i})$ above, we now analyze its behavior when applied to test data distributions that differ from the trainset, denoted as P_{test} versus P_{train} . Recall that any joint distribution over images and text can be factored into a product over a language prior and an image likelihood $P(\mathbf{t}, \mathbf{i}) = P(\mathbf{t})P(\mathbf{i}|\mathbf{t})$. Our analysis makes the strong assumption that the image likelihood $P(\mathbf{i}|\mathbf{t})$ is identical across the train and test data, but the language prior $P(\mathbf{t})$ may differ. Intuitively, this assumes that the visual appearance of entities (such as a "white duck") remains consistent across the training and test data, but the

frequency of those entities (as manifested in the set of captions $P(\mathbf{t})$) may vary. We can now derive $P_{test}(\mathbf{t}|\mathbf{i})$ via Bayes rule:

$$P_{test}(\mathbf{t}|\mathbf{i}) \propto P(\mathbf{i}|\mathbf{t})P_{test}(\mathbf{t}) \quad (4.2)$$

$$= P(\mathbf{i}|\mathbf{t})\frac{P_{train}(\mathbf{t})}{P_{train}(\mathbf{t})}P_{test}(\mathbf{t}) \quad (4.3)$$

$$\propto P_{train}(\mathbf{t}|\mathbf{i})\frac{P_{test}(\mathbf{t})}{P_{train}(\mathbf{t})} \quad (4.4)$$

The above shows that the generative pre-training score $P_{train}(\mathbf{t}|\mathbf{i})$ need simply be weighted by the *ratio* of the language priors in the testset versus trainset. Intuitively, if a particular text caption appears *more* often in the testset than the trainset, one should *increase* the score reported by the generative model. However, one often does not have access to the text distribution on the testset. For example, real-world deployments and benchmark protocols may not reveal this. In such cases, one can make two practical assumptions; either the language distribution on test is identical to train, or it is uninformative/uniform (see [Figure 4.1](#)):

Scenario 1:

$$P_{test}(\mathbf{t}) = P_{train}(\mathbf{t}) \quad \Rightarrow \quad \text{Optimal score is } P_{train}(\mathbf{t}|\mathbf{i}) \quad (4.5)$$

Scenario 2:

$$P_{test}(\mathbf{t}) \text{ is uniform.} \quad \Rightarrow \quad \text{Optimal score is } \frac{P_{train}(\mathbf{t}|\mathbf{i})}{P_{train}(\mathbf{t})} \quad (4.6)$$

Tunable α . In reality, a testset might be a mix of both scenarios. To model this, we consider a soft combination where the language prior on the testset is assumed to be a flattened version of the language prior on the trainset, for some temperature parameter $\alpha \in [0, 1]$:

$$P_{test}(\mathbf{t}) \propto P_{train}(\mathbf{t})^{1-\alpha} \quad \Rightarrow \quad \text{Optimal score is } \frac{P_{train}(\mathbf{t}|\mathbf{i})}{P_{train}(\mathbf{t})^\alpha} \quad (4.7)$$

By setting α to 0 or 1, one can obtain the two scenarios described above. Some deployments (or benchmarks) may benefit from tuning α on a held-out validation set.

Implications for retrieval benchmarks. We speculate some benchmarks like ARO-Flickr [208] are close to scenario 1 because they include negative captions that are *implausible*, such as “a white duck the its wings while in water spreads”. Such captions will have a low score under the language prior $P_{train}(\mathbf{t})$ and so reporting the raw generative score $P_{train}(\mathbf{t}|\mathbf{i})$ (that keeps its language prior or bias) will improve accuracy. In fact, we show that applying a *blind* language model (that ignores all image evidence) can itself often identify the correct caption. On the other hand, for test datasets with more *realistic* negative captions (scenario 2), it may be useful to remove the language bias of the trainset, since that will prefer to match to common captions (even if they do not necessarily agree with the input image). This appears to be the case for SugarCrepe [74], which uses LLMs like ChatGPT to ensure that the negative captions are realistic.

Relationship to prior approaches. Our approach to debiasing is reminiscent of mutual information, which can also be seen as a method for removing the effect of marginal priors when computing joint probability scores. In fact, our [section 4.6](#) derives that α -debiasing is equivalent to a form of pointwise mutual information (PMI) known as PMI^k for $k = \frac{1}{\alpha}$.

4.4 Experimental results on I-to-T retrieval

In this section, we verify our hypothesis on I-to-T retrieval benchmarks using state-of-the-art multimodal generative VLMs. In particular, we adopt image-conditioned language models such as BLIP [103] as the learned estimator of $P_{train}(\mathbf{t}|\mathbf{i})$. Then, we discuss how we perform Monte Carlo estimation of $P_{train}(\mathbf{t})$, including a novel efficient sampling method based on “content-free” Gaussian noise images. Finally, we show the state-of-the-art results of our generative approach on existing I-to-T retrieval tasks.

Preliminaries. We leverage OTS image-conditioned language models [3, 105, 206] to estimate $P_{train}(\mathbf{t})$. For ablation, we use the open-sourced BLIP models [103], trained on public image-text corpora using discriminative (ITC and ITM) and generative (captioning) objectives. Discriminative objectives typically model $P(\text{match}|\mathbf{t}, \mathbf{i})$. For example, ITCScore calculates cosine similarity scores between image and text features using a dual-encoder; ITMScore jointly embeds image-text pairs via a fusion-encoder

4. Revisiting the Role of Language Priors in Vision-Language Models

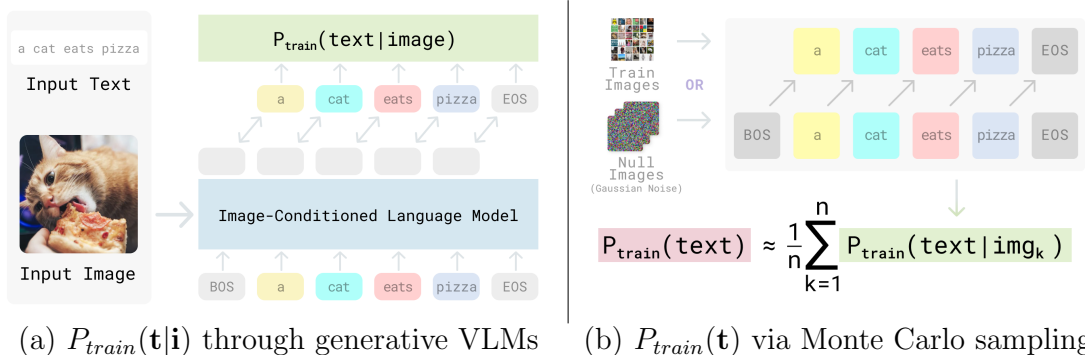


Figure 4.2: **Estimating $P_{train}(\mathbf{t}|\mathbf{i})$ and $P_{train}(\mathbf{t})$ from generative VLMs.** Figure (a) shows how image-conditioned language models such as Li et al. [103] that generate text based on an image can be repurposed for computing $P_{train}(\mathbf{t}|\mathbf{i})$, which is factorized as a product of $\prod_{k=1}^m P(t_k|t_{<k}, \mathbf{i})$ for a sequence of m tokens. These terms can be efficiently computed in *parallel*, unlike *sequential* token-by-token prediction for text generation. Figure (b) shows two approaches for Monte Carlo sampling of $P_{train}(\mathbf{t})$. While the straightforward approach is to sample trainset images, we find that using as few as three “null” (Gaussian noise) images can achieve more robust estimates.

and returns softmax scores from a binary classifier. Lastly, we term the generative score as **Visual Generative Pre-Training Score (VisualGPTScore)**. While BLIP is pre-trained using all three objectives, this generative score has not been applied to discriminative tasks before our work.

Implementing VisualGPTScore. Our method calculates an average of the log-likelihoods of t_k at each token position k and applies an exponent to cancel the log:

$$\text{VisualGPTScore}(\mathbf{t}, \mathbf{i}) := e^{\frac{1}{m} \sum_{k=1}^m \log(P(t_k|t_{<k}, \mathbf{i}))} \quad (4.8)$$

To condition on an input image, BLIP uses a multimodal casual self-attention mask [103] in its image-grounded text decoder, i.e., each text token attends to all its preceding vision and text tokens. We emphasize that VisualGPTScore has the same computational cost as ITMScore, which uses the same underlying transformer but with a bi-directional self-attention mask to encode an image-text pair. We address potential biases of this estimator in [section 4.8](#).

Estimating $P_{train}(\mathbf{t})$ using Monte Carlo sampling (oracle approach). Given $P_{train}(\mathbf{t}|\mathbf{i})$, we can estimate $P_{train}(\mathbf{t})$ via classic Monte Carlo sampling [168], by

drawing n images from the train distribution, such as LAION114M [165] for BLIP:

$$P_{train}(\mathbf{t}) \approx \frac{1}{n} \sum_{k=1}^n P_{train}(\mathbf{t}|\mathbf{i}_k) \quad (4.9)$$

Reducing sampling cost with content-free images (our approach). The above Equation 4.9 requires many trainset samples to achieve robust estimates. To address this, we draw inspiration from [217], which uses a *content-free* text prompt “N/A” to calibrate the probability of a text from LLMs, i.e., $P(\mathbf{t}|\text{“N/A”})$. To apply this to our generative VLMs, we choose to sample “null” inputs as Gaussian noise images. As a result, our approach requires as few as three images to compute Eq. 4.9 by sampling from Gaussian noise images with a mean of 0.4 and a standard deviation of 0.25. We find this method to be less computationally demanding and just as effective as sampling thousands of images from trainset. We provide a visual illustration of this method in Figure 4.2-b. We include sampling details in section 4.7.

Benchmarks and evaluation protocols. We comprehensively report on four popular I-to-T retrieval benchmarks, including ARO [208], Crepe [123], SugarCrepe [74], and VL-CheckList [216]. In these datasets, each image has a single positive caption and multiple negative captions. ARO [208] has four datasets: VG-Relation, VG-Attribution, COCO-Order, and Flickr30k-Order. SugarCrepe [74] has three datasets: Replace, Swap, and Add. For Crepe [123], we use the entire productivity set and report on three datasets: Atom, Negate, and Swap. VL-CheckList [216] has three datasets: Object, Attribute, and Relation.

SOTA performance on all four benchmarks. In Table 4.1, we show that our OTS generative approaches, based on the BLIP model pre-trained on LAION-114M with ViT-L image encoder, achieves state-of-the-art results on all benchmarks. We outperform the best discriminative VLMs, including LAION5B-CLIP, and consistently surpass other heavily-engineered solutions, including NegCLIP, SyViC, MosaiCLIP, DAC, SVLC, SGVL, Structure-CLIP, all of which fine-tune CLIP on much more data. For reference, we also include results of text-only Vera and Grammar from Hsieh et al. [74]. To show that even the most recent SugarCrepe is not exempt from language biases, we run two more text-only methods:

1. $P_{LLM}(\mathbf{t})$: passing captions into a pure LLM, such as BART-base [207], FLAN-

T5-XL [25], and OPT-2.7B [215], to compute a text-only GPTScore [43].

2. $P_{train}(\mathbf{t})$: passing both captions and Gaussian noise images to BLIP as shown in Figure 4.2.

Visualization of α -debiasing. Finally, we observe that α -debiasing can consistently improve the performance. For visualization, we attach the results of α -debiasing in Table 4.2. We show side-by-side frequency charts of $P_{train}(\mathbf{t})$ for positive and negative captions.

4.5 Additional Experimental Results

In this section, we apply our OTS generative approaches to more benchmarks, including two compositionality benchmarks Winoground [182] and EqBen [187], two classic large-scale retrieval benchmarks COCO [112] and Flickr30K [205], and zero-shot image classification on ImageNet [30]. While naively applying VisualGPTScore leads to bad performance on these benchmarks, our training-free debiasing solution can consistently improve its performance with a held-out validation set. Furthermore, we derive the optimal text-to-image (T-to-I) retrieval objective and show that OTS generative scores can achieve robust T-to-I performance.

Evaluation protocols of Thrush et al. [182]. While prior analysis [36, 208] suggests that Winoground is too out-of-distribution to evaluate compositionality, we argue that evaluation protocols of Winoground and EqBen are more robust for future evaluations of VLMs. In these two benchmarks, each sample consists of two image-text pairs, ensuring **uniform image and text priors**. For simplicity, we consider a single Winoground sample: $(\mathbf{i}_0, \mathbf{t}_0)$ and $(\mathbf{i}_1, \mathbf{t}_1)$. The joint probabilities are $P_{test}(\mathbf{i}_0, \mathbf{t}_0) = P_{test}(\mathbf{i}_1, \mathbf{t}_1) = 0.5$. Meanwhile, $P_{test}(\mathbf{i}_0, \mathbf{t}_1) = P_{test}(\mathbf{i}_1, \mathbf{t}_0) = 0$. Applying the law of total probability gives $P_{test}(t_0) = P_{test}(t_1) = 0.5$. A similar derivation can show that image priors are uniform too. In addition, Winoground’s evaluation metrics (text score and image score) penalize unimodal shortcut solutions. For example, in I-to-T retrieval, the *text score* gets 1 point only if *both images* are matched to the correct caption. Therefore, “blind” solutions that choose the same text regardless of images will get 0 text score. Similarly, for T-to-I retrieval, the *image score* gets 1 point only if *both captions* are matched to the correct image.

Tuning α through cross validation. In [Table 4.3-a](#), we first show that OTS generative scores without debiasing ($\alpha=0$) lead to inferior performance on these I-to-T benchmarks. This confirms the importance of α -debiasing; even a simple $\alpha = 1$ can consistently and often significantly improve their I-to-T results. Furthermore, we try to use a held-out validation set to tune for optimal $\alpha \in [0, 1]$. We sample half of the data as validation set to search for α_{val}^* (using a step size of 0.001) and report the performance on the other half. We repeat this process 10 times to and report the mean and std. We observe that the optimal alpha is usually stable under the same dataset, regardless of the sampled val set. For COCO and Flickr30K, we perform α -debiasing using Recall@1 (R@1) on the official validation split. Because sampling additional Gaussian noise images can be too costly on these large-scale benchmarks, we directly approximate $P_{train}(\mathbf{t})$ by averaging the scores of testset images, without incurring any computational cost. More ablation studies such as α -debiasing using testset can be found in [section 4.7](#). We also include the results of the ITMScore of BLIP for reference. While our debiasing solution can always boost performance, we observe that generative approaches still lag behind the ITMScore for these two retrieval benchmarks. This motivates us to study biases of generative scores towards more “common” texts in [section 4.8](#). Finally, we report the zero-shot classification accuracy on ImageNet1K, which can be viewed as an image-to-text retrieval task that selects the most fit textual label (out of 1000) for each image. For cross validation results on ImageNet, we simply use one-shot trainset samples from [114].

Extending to T-to-I retrieval. Though not the focus of our work, we also show that image-conditioned language models can be applied to T-to-I retrieval. Given a text caption \mathbf{t} , we can rewrite the Bayes optimal T-to-I retrieval objective as:

$$P_{test}(\mathbf{i}|\mathbf{t}) \propto P_{train}(\mathbf{t}|\mathbf{i}) * P_{train}(\mathbf{i}) \quad (4.10)$$

[Equation 4.10](#) is hard to implement because we do not have access to $P_{train}(\mathbf{i})$. However, when $P_{train}(\mathbf{i})$ is approximately uniform, one can directly apply $P_{train}(\mathbf{t}|\mathbf{i})$ for optimal performance. We report T-to-I performance on all four benchmarks in [Table 4.3-b](#), where our generative approach obtain competitive results compared against ITMScore, presumably because T-to-I retrieval is less affected by language biases.

Results with BLIP-2 [105]. In Table 4.4, we show that our α -debiasing approach generalize to the SOTA captioning model BLIP-2. BLIP-2 leverages powerful frozen pre-trained image encoders [40] and large language models [25, 215] to bootstrap vision-language pre-training. It proposes a lightweight Querying Transformer (Q-Former) that is trained in two stages. Similar to BLIP [103], Q-Former is a mixture-of-expert model that can calculate ITC, ITM, and captioning loss given an image-text pair. Additionally, it introduces a set of trainable query tokens, whose outputs serve as *visual soft prompts* prepended as inputs to LLMs. In its first training stage, Q-Former is fine-tuned on the same LAION dataset using the same objectives (ITC+ITM+captioning) as BLIP. In the second stage, the output query tokens from Q-Former are fed into a frozen language model, such as FLAN-T5 [25], after a linear projection trained only with captioning loss. We report results for both the first-stage model (denoted as Q-Former) and the second-stage model which employs FLAN-T5 [25] as the frozen LLM.

4.6 Comparison to PMI^k

By assuming $P_{test}(\mathbf{t})$ to be a “flatten” version of $P_{train}(\mathbf{t})$, our Equation 4.7 can interpolate between scenario 1 (same train and test priors) and 2 (balanced test priors):

$$P_{test}(\mathbf{t}) \propto P_{train}(\mathbf{t})^{1-\alpha} \quad \Rightarrow \text{Optimal score is } \frac{P_{train}(\mathbf{t}|\mathbf{i})}{P_{train}(\mathbf{t})^\alpha} \quad (4.11)$$

In fact, the above equation can be rewritten using the language of PMI^k [29, 158], a well-known variant of PMI that controls the amount of debiasing [100, 101, 191] in information retrieval:

$$\frac{P_{train}(\mathbf{t}|\mathbf{i})}{P_{train}(\mathbf{t})^\alpha} = \frac{P_{train}(\mathbf{t}, \mathbf{i})}{P_{train}(\mathbf{i})P_{train}(\mathbf{t})^\alpha} \quad (4.12)$$

$$\propto \frac{P_{train}(\mathbf{t}, \mathbf{i})^\frac{1}{\alpha}}{P_{train}(\mathbf{i})P_{train}(\mathbf{t})} \quad , \text{ as } P_{train}(\mathbf{i}) \quad (4.13)$$

$$= \text{pmi}_{P_{train}}^k(\mathbf{t}, \mathbf{i}), \text{ where } k = \frac{1}{\alpha} \geq 1 \quad (4.14)$$

where

$$\text{pmi}_P(\mathbf{t}, \mathbf{i}) = \frac{P(\mathbf{t}, \mathbf{i})}{P(\mathbf{t})P(\mathbf{i})} = \frac{P(\mathbf{t}|\mathbf{i})}{P(\mathbf{t})} = \frac{P(\mathbf{i}|\mathbf{t})}{P(\mathbf{i})} \quad (4.15)$$

PMI is an information-theoretic measure that quantifies the *association* between two variables [65, 172, 204]. In the context of image-text retrieval, it measures how much more (or less) likely the image-text pair co-occurs than if the two were independent. Eq. 4.15 has found applications in diverse sequence-to-sequence modelling tasks [100, 101, 191] as a retrieval (reranking) objective. Compared to the conditional likelihood $P(\mathbf{t}|\mathbf{i})$, PMI reduces the learned bias for preferring "common" texts with high marginal probabilities $P(\mathbf{t})$ [100, 101, 191]. This can be an alternative explanation for the effectiveness of our debiasing solutions.

4.7 Ablation Studies on α -Debiasing

Estimating $P_{train}(\mathbf{t})$ via null (Gaussian noise) images is more sample-efficient.

We use Winoground to show that sampling Gaussian noise images to calculate $P_{train}(\mathbf{t})$ can be more efficient than sampling trainset images. As demonstrated in Table 4.5, a limited number of Gaussian noise images (e.g., 3 or 10) can surpass the results obtained with 1000 LAION images. Moreover, using null images produces less variance in the results.

Details of Gaussian noise samples. Unless otherwise specified, the Gaussian noise images are sampled with a mean of 1.0 and a standard deviation of 0.25. By default, we use 100 images for Winoground, 30 images for EqBen, 1 image for ImageNet, and 3 images for the rest of the benchmarks. We leave more advanced techniques of generating null images to future work.

Alternative approach on COCO/Flickr30k: estimating $P_{train}(\mathbf{t})$ using testset images.

For large-scale retrieval benchmarks like COCO [112] and Flickr30k [205], we can directly average scores of all candidate images (in the order of thousands) to efficiently approximate $P_{train}(\mathbf{t})$ without the need to sample additional images. This approach incurs zero computation cost as we have already pre-computed scores between each candidate image and text. We show in Table 4.6 that using testset images indeed results in better performance than sampling 3 Gaussian noise

images.

Tuning α with a validation set. In Table 4.7, similar performance trends are observed across validation and test splits of COCO and Flickr30k I-to-T retrieval benchmarks using the same $\alpha \in [0, 1]$. Furthermore, α_{test}^* and α_{val}^* are empirically close. As such, our method can function as a reliable training-free debiasing method. Future studies may explore fine-tuning methods to further improve the debiasing performance.

4.8 Is VisualGPTScore a Biased Estimator?

Retrieval performance on trainset (LAION). This paper is built on the assumption that VisualGPTScore is a reliable estimator of $P_{train}(\mathbf{t}|\mathbf{i})$. However, this simplifying assumption does not completely hold for the BLIP model we examine. We speculate that such OTS generative scores are biased towards more common texts. We witness this same phenomenon in Table 4.8, where we perform image-text retrieval on random subsets from training distribution LAION-114M [103].

Modelling the language bias in VisualGPTScore. As evidenced in Table 4.8, we believe VisualGPTScore is biased towards more common texts due to modelling error. To consider this error in our analysis, we rewrite the VisualGPTScore as:

$$\mathbf{VisualGPTScore}(\mathbf{t}, \mathbf{i}) := \hat{P}_{train}(\mathbf{t}|\mathbf{i}) = P_{train}(\mathbf{t}|\mathbf{i}) \cdot P_{train}(\mathbf{t})^\beta \quad (4.16)$$

where \hat{P} represents the (biased) model estimate and P represents the true distribution. The model bias towards common texts is encoded by an unknown parameter β .

Monte Carlo estimation using \hat{P} . Because our Monte Carlo sampling method relies on $\hat{P}_{train}(\mathbf{t}|\mathbf{i})$, it is also a biased estimator of $P_{train}(\mathbf{t})$:

$$\hat{P}_{train}(\mathbf{t}) := \frac{1}{n} \sum_{k=1}^n \hat{P}_{train}(\mathbf{t}|\mathbf{i}_k) = P_{train}(\mathbf{t})^{1+\beta}. \quad (4.17)$$

Rewriting optimal I-to-T objective with \hat{P} . We can rewrite Equation 4.4 as:

$$P_{test}(\mathbf{t}|\mathbf{i}) \propto P_{train}(\mathbf{t}|\mathbf{i}) \frac{P_{test}(\mathbf{t})}{P_{train}(\mathbf{t})} \quad (4.18)$$

$$= \hat{P}_{train}(\mathbf{t}|\mathbf{i}) \frac{P_{test}(\mathbf{t})}{P_{train}(\mathbf{t})^{1+\beta}} \quad (4.19)$$

$$= \hat{P}_{train}(\mathbf{t}|\mathbf{i}) \frac{P_{test}(\mathbf{t})}{\hat{P}_{train}(\mathbf{t})} \quad (4.20)$$

α -debiasing with \hat{P} . Using Equation 4.20, we can reformulate α -debiasing (Equation 4.7) as follows:

$$P_{test}(\mathbf{t}) \propto P_{train}(\mathbf{t})^{1-\alpha} \quad \Rightarrow \text{Optimal score is } \frac{\hat{P}_{train}(\mathbf{t}|\mathbf{i})}{\hat{P}_{train}(\mathbf{t})^\alpha} \quad (4.21)$$

where $\alpha = \frac{\hat{\alpha}+\beta}{1+\beta}$. Notably, the above equation has the same structure as before (Equation 4.7). This implies that even if $P_{train}(\mathbf{t}) = P_{test}(\mathbf{t})$, we still anticipate $\alpha = \frac{\beta}{1+\beta} \neq 0$. This accounts for why the optimal α is not 0 when we perform I-to-T retrieval on trainset in Table 4.8.

Implication for vision-language modelling. Our analysis indicates that similar to generative LLMs [100, 101], contemporary image-conditioned language models also experience issues related to imbalanced learning [89]. Potential solutions could be: (a) refined sampling techniques for Monte Carlo estimation of $P(\mathbf{t})$ such as through dataset distillation [197], and (b) less biased modelling of $P(\mathbf{t}|\mathbf{i})$ such as through controllable generation [90].

4.9 Discussion and Limitations

Summary. Our study shows the efficacy of *generative* pre-training scores in solving *discriminative* tasks. With the rise of generative pre-training in recent models like GPT-4 [139], we see our work as a reliable starting point for future tasks. We present a first-principles analysis to account for mismatching distributions over text between train and test data. Based on this, we introduce a robust training-free (zero-shot) solution to debias linguistic priors in generative scores, achieving consistent and often

significant improvement on all I-to-T retrieval tasks. Our thorough analysis also explains the performance discrepancy of generative scores on different benchmarks, and we hope it can encourage future work to revisit the issue of language biases in vision-language benchmarks.

Limitations and future work. Our approach depends on generative VLMs pre-trained on noisy web datasets, which may result in inherited biases [127]. We do not explore fine-tuning techniques due to computational constraints, but it is possible to improve the I-to-T retrieval performance using hard negative samples, such as with controllable generation [90]. Furthermore, our analysis is based on simplified assumptions. For instance, the image-conditioned language model might not accurately represent $P_{train}(\mathbf{t}|\mathbf{i})$, a phenomenon we examine in [section 4.8](#). Estimating $P_{train}(\mathbf{t})$ by sampling Gaussian noise images can be suboptimal; future VLMs could directly model $P_{train}(\mathbf{t})$, or use techniques like coreset selection [52] or dataset distillation [197] to sample more representative images. While VisualGPTScore shows competitive performance, it still suffers from a high inference cost compared to ITCScore especially for large-scale retrieval tasks; therefore, distilling it into dual-encoder head [130] can be a promising future direction. Finally, we leave debiasing on the T-to-I retrieval task for future work.

4. Revisiting the Role of Language Priors in Vision-Language Models

Score	Method	ARO				
		Rel	Attr	COCO	Flickr	
Random	-	50.0	50.0	20.0	20.0	
Text-Only	Vera	61.7	82.6	59.8	63.5	
	Grammar	59.6	58.4	74.3	76.3	
$P_{LLM}(t)$	BART	81.1	73.6	95.0	95.2	
	Flan-T5	84.4	76.5	98.0	98.2	
	OPT	84.7	79.8	97.9	98.6	
$P_{train}(t)$	BLIP	87.6	80.7	98.6	99.1	
$P(\text{match} t, i)$	CLIP	59.0	62.0	59.0	46.0	
	LAION2B-CLIP	51.6	61.9	25.2	30.2	
	LAION5B-CLIP	46.1	57.8	26.1	31.0	
	NegCLIP	81.0	71.0	91.0	86.0	
	Structure-CLIP	83.5	85.1	-	-	
	SyViC	80.8	72.4	92.4	87.2	
	SGVL	-	-	87.2	91.0	
	MosaiCLIP	82.6	78.0	87.9	86.3	
	DAC-LLM	81.3	73.9	94.5	95.7	
	DAC-SAM	77.2	70.5	91.2	93.9	
	BLIP-ITC	63.1	81.6	34.3	41.7	
	BLIP-ITM	58.7	90.3	45.1	51.3	
	$P_{train}(t i)$	Ours ($\alpha = 0$)	89.1	95.3	99.4	99.5
	$P_{train}(t)^\alpha$	Ours ($\alpha = 1$)	68.1	87.9	32.4	44.5
	Ours ($\alpha = \alpha^*$)	89.1	95.4	99.4	99.5	

(a) Accuracy on ARO

Score	Method	SugarCrepe		
		Replace	Swap	Add
Random	-	50.0	50.0	50.0
Text-Only	Vera	49.5	49.3	49.5
	Grammar	50.0	50.0	50.0
$P_{LLM}(t)$	BART	48.4	51.9	61.2
	Flan-T5	51.4	57.6	40.9
	OPT	58.5	66.6	45.8
$P_{train}(t)$	BLIP	75.9	77.1	70.9
$P(\text{match} t, i)$	CLIP	80.8	63.3	75.1
	LAION2B-CLIP	86.5	68.6	88.4
	LAION5B-CLIP	85.0	68.0	89.6
	NegCLIP	88.3	76.2	90.2
	BLIP-ITC	85.8	73.8	85.7
	BLIP-ITM	88.7	81.3	87.6
	Ours ($\alpha = 0$)	93.3	91.0	91.0
$P_{train}(t i)$	Ours ($\alpha = 1$)	83.2	85.5	85.9
$P_{train}(t)^\alpha$	Ours ($\alpha = \alpha^*$)	95.1	92.4	87.4

(c) Accuracy on SugarCrepe

Score	Method	VL-CheckList			
		Object	Attribute	Relation	
Random	-	50.0	50.0	50.0	
Text-Only	Vera	82.5	74.0	85.7	
	Grammar	58.0	52.4	68.5	
$P_{LLM}(t)$	BART	52.0	51.0	45.1	
	Flan-T5	60.3	55.0	49.3	
	OPT	59.3	48.8	60.0	
$P_{train}(t)$	BLIP	68.2	58.7	75.9	
$P(\text{match} t, i)$	CLIP	81.6	67.6	63.1	
	LAION2B-CLIP	84.7	67.8	66.5	
	LAION5B-CLIP	87.9	70.3	63.9	
	NegCLIP	81.4	72.2	63.5	
	SyViC	-	70.4	69.4	
	SGVL	85.2	78.2	80.4	
	SLVC	85.0	72.0	69.0	
	DAC-LLM	87.3	77.3	86.4	
	DAC-SAM	88.5	75.8	89.8	
	BLIP-ITC	90.6	80.3	73.5	
	BLIP-ITM	89.9	80.7	67.7	
	$P_{train}(t i)$	Ours ($\alpha = 0$)	92.6	78.7	90.8
	$P_{train}(t)^\alpha$	Ours ($\alpha = 1$)	90.4	77.6	77.8
		Ours ($\alpha = \alpha^*$)	94.4	82.1	92.8

(b) Accuracy on VL-CheckList

Score	Method	Crepe		
		Atom	Swap	Negate
Random	-	16.7	16.7	16.7
Text-Only	Vera	43.7	70.8	66.2
	Grammar	18.2	50.9	9.8
$P_{LLM}(t)$	BART	38.8	53.3	44.4
	Flan-T5	43.0	69.5	13.6
	OPT	53.3	72.7	5.0
$P_{train}(t)$	BLIP	55.4	69.7	60.8
$P(\text{match} t, i)$	CLIP	22.3	26.6	28.8
	LAION2B-CLIP	23.6	24.8	18.0
	LAION5B-CLIP	24.2	23.9	20.1
	BLIP-ITC	24.8	17.7	26.5
	BLIP-ITM	29.5	20.7	25.5
	Ours ($\alpha = 0$)	73.2	78.1	79.6
	$P_{train}(t i)$	Ours ($\alpha = 1$)	20.6	28.3
$P_{train}(t)^\alpha$	Ours ($\alpha = \alpha^*$)	73.3	78.1	79.6

(d) Accuracy on Crepe

Table 4.1: **OTS generative VLMs are SOTA on image-to-text retrieval benchmarks.** We begin by evaluating blind language models (in red). Surprisingly, this already produces SOTA accuracy on certain benchmarks such as ARO-Flickr, compared to the best discriminative approaches (in gray). We also find that blind inference of generative VLMs, $P_{train}(t)$ via sampling Gaussian noise images (in blue), often performs better and achieve above-chance performance even on the most recent SugarCrepe. Next, we show that simply repurposing a generative VLM’s language generation head for computing image-text scores (VisualGPTScore in yellow), which corresponds to $\alpha = 0$, consistently produces SOTA accuracy across all benchmarks. Finally, debiasing this score by tuning α on val set (in green) further improves performance, establishing the new SOTA.

4. Revisiting the Role of Language Priors in Vision-Language Models

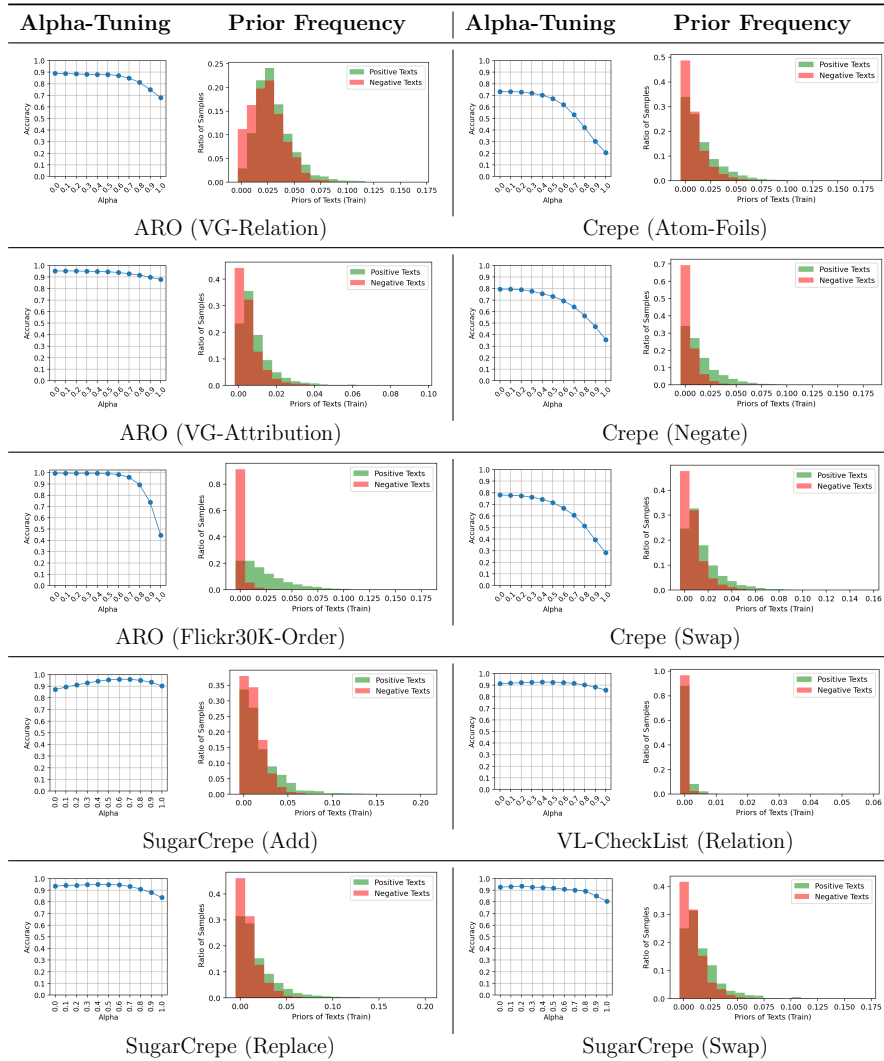


Table 4.2: α -debiasing on I-to-T benchmarks and $P_{train}(\mathbf{t})$ frequency charts of both positive and negative captions. Increasing α from 0 to 1 hurts performance on benchmarks with non-sensical negative captions such as ARO and Crepe. Such negative captions are easier to identify because of their low score under the language prior $P_{train}(\mathbf{t})$, implying such benchmarks may even be solved with blind algorithms that avoid looking at images. On the other hand, for benchmarks like SugarCrepe with more balanced $P_{train}(\mathbf{t})$ between positives and negatives, tuning α may lead to performance gain.

4. Revisiting the Role of Language Priors in Vision-Language Models

Metric	Benchmark	ITMScore	$\frac{P_{train}(\mathbf{t} \mathbf{i})}{P_{train}(\mathbf{t})^\alpha}$				α_{val}^*	Metric	Benchmark	ITMScore	$P_{train}(\mathbf{t} \mathbf{i})$
			$\alpha=0$	$\alpha=1$	$\alpha=\alpha_{val}^*$	α_{val}^*					
Text Score	Winoground	35.5 _(2.4)	27.5 _(2.3)	33.7 _(2.4)	36.6 _(2.6)	0.855 _(0.023)	Image Score	Winoground	15.8	21.5	
	EqBen	26.1 _(0.3)	9.6 _(0.2)	19.8 _(0.3)	19.8 _(0.3)	0.992 _(0.007)		EqBen	20.3	26.1	
R@1 / R@5	COCO	71.9 / 90.6	19.7 / 40.6	46.2 / 73.1	48.0 / 74.2	0.819	R@1 / R@5	COCO	54.8 / 79.0	55.6 / 79.2	
	Flickr30k	88.8 / 98.2	34.6 / 59.0	58.7 / 88.0	63.6 / 89.2	0.719		Flickr30k	77.8 / 93.9	76.8 / 93.4	
Accuracy	ImageNet1K	37.4	18.6	36.2	40.0	0.670	(b) T-to-I retrieval				

(a) α -debiasing on val sets for I-to-T retrieval

(b) T-to-I retrieval

Table 4.3: **Additional results on Winoground/EqBen/COCO/Flickr30K/ImageNet1K.**

Table (a) shows the importance of α -debiasing on these compositionality and large-scale retrieval benchmarks. While OTS generative scores do not work well, debiasing with a larger α close to 1 can consistently and often significantly improve I-to-T performance. To highlight the improvement, we mark results without debiasing ($\alpha = 0$) (in yellow), debiasing with a fixed $\alpha = 1$ (in pink), and cross-validation using held-out val sets ($\alpha = \alpha_{val}^*$) (in green). Table (b) shows that OTS generative scores can obtain favorable results on all T-to-I retrieval tasks, competitive with the ITMScore.

Benchmark	Model	$\frac{P_{train}(\mathbf{t} \mathbf{i})}{P_{train}(\mathbf{t})^\alpha}$			
		$\alpha=0$	$\alpha=1$	$\alpha=\alpha^*$	α^*
Winoground	BLIP	27.0	33.0	36.5	0.836
	BLIP2-QFormer	24.3	29.3	33.0	0.882
	BLIP2-FlanT5	25.3	31.5	34.3	0.764
EqBen (Val)	BLIP	9.6	19.8	19.8	0.982
	BLIP2-QFormer	12.2	21.9	22.2	0.969
	BLIP2-FlanT5	8.5	22.0	22.0	1.000

Table 4.4: **α -debiasing consistently improves BLIP-2 on balanced VL benchmarks.**

We show that α -debiasing, even with a fixed $\alpha=1$, can consistently improve BLIP-2 performance on challenging Winoground and EqBen.

Sample Size	Gaussian Noise Images		Trainset Images	
	$\alpha=\alpha_{test}^*$	α_{test}^*	$\alpha=\alpha_{test}^*$	α_{test}^*
3	35.95 _(0.5)	0.821 _(0.012)	32.20 _(1.6)	0.706 _(0.150)
10	36.25 _(0.4)	0.827 _(0.016)	33.60 _(0.9)	0.910 _(0.104)
100	36.35 _(0.1)	0.840 _(0.010)	34.70 _(0.6)	0.910 _(0.039)
1000	36.25 _(0.0)	0.850 _(0.000)	35.15 _(0.3)	0.960 _(0.033)

Table 4.5: **Comparing sampling of Gaussian noise images and trainset images for estimating $P_{train}(\mathbf{t})$.** We report text scores of α -debiasing on Winoground I-to-T retrieval task. We ablate 3/10/100/1000 Gaussian noise and LAION samples and report both mean and std using 5 sampling seeds. The optimal $\alpha^* \in [0, 1]$ is searched on testset via a step size of 0.001. The Gaussian noise images are sampled with a mean calculated from the LAION subset and a fixed std of 0.25.

4. Revisiting the Role of Language Priors in Vision-Language Models

Metric	Benchmark	$P_{train}(\mathbf{t} \mathbf{i})$	Sampling Method	$\frac{P_{train}(\mathbf{t} \mathbf{i})}{P_{train}(\mathbf{t})^\alpha}$		
				$\alpha=1$	$\alpha=\alpha_{val}^*$	α_{val}^*
R@1 / R@5	COCO	19.7 / 40.6	Testset Images	46.2 / 73.1	48.0 / 74.2	0.819
			Null Images	24.4 / 52.6	40.4 / 66.6	0.600
	Flickr30k	34.6 / 59.0	Testset Images	58.7 / 88.0	63.6 / 89.2	0.719
			Null Images	27.8 / 62.2	48.5 / 79.0	0.427

Table 4.6: **I-to-T retrieval on COCO/Flickr30k using different sampling methods.** Estimating $P_{train}(\mathbf{t})$ by averaging the scores of testset images (with zero computational cost) demonstrates superior performance compared to sampling additional Gaussian noise images.

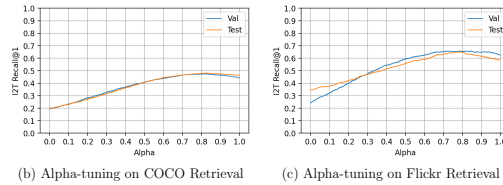
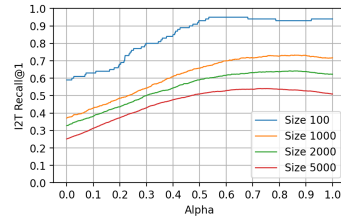


Table 4.7: **α -debiasing results on both val set and test set for COCO/Flickr30k I-to-T retrieval.** We observe that validation and test performance are strongly correlated while we interpolate $\alpha \in [0, 1]$.

Dataset Size	I-to-T Retrieval				T-to-I Retrieval		
	ITM	$\frac{P_{train}(\mathbf{t} \mathbf{i})}{P_{train}(\mathbf{t})^\alpha}$			ITM	$P_{train}(\mathbf{t} \mathbf{i})$	
		$\alpha=0$	$\alpha=1$	$\alpha=\alpha^*$			α^*
100	96.0	59.0	94.0	95.0	0.535	95.0	97.0
1000	90.9	37.1	71.7	85.7	0.733	92.0	93.1
2000	87.2	32.8	62.3	64.3	0.840	87.8	89.8
5000	79.8	25.1	50.9	54.1	0.727	81.9	84.4

(a) Performance on LAION trainset retrieval



(b) Alpha-tuning on LAION

Table 4.8: **Retrieval performance on randomly sampled LAION114M subsets with varied sizes.** Table (a) shows that while OTS generative scores are robust for T-to-I retrieval, its performance degrades on I-to-T retrieval tasks when the number of candidate texts increases. This implies that OTS generative scores suffer from language biases towards certain texts even in the training set. Nonetheless, we show that our debiasing solution using either $\alpha = 1$ or optimal $\alpha^* \in [0, 1]$ with a step size of 0.001, can consistently boost the performance. Figure (b) visualizes α -debiasing results on LAION subsets, where each curve represents a different sample size.

4. *Revisiting the Role of Language Priors in Vision-Language Models*

Chapter 5

Conclusions

This thesis has introduced novel advancements in vision-language models (VLMs), significantly enhancing their alignment with user objectives and broadening their application scope. The "cross-modal adaptation" method represents a major stride in visual classification, utilizing minimal data and integrating textual and audio inputs to refine VLMs. Additionally, the novel black-box approach leveraging large language models (LLMs) like ChatGPT has shown remarkable improvements in one-shot visual classification and text-to-image tasks, demonstrating the potential of synergizing VLMs and LLMs.

However, despite these advancements, the thesis also highlights the inherent limitations of current VLMs in handling complex compositional reasoning. The strategies developed to assess and improve the models' abilities in managing detailed object compositions, attributes, and relationships underscore the ongoing need for research in this area. This work not only contributes to a deeper understanding of the capabilities and limitations of VLMs but also lays a foundation for future research aimed at refining these models for more nuanced and sophisticated tasks.

In summary, the research presented in this thesis not only pushes the boundaries of current vision-language modeling but also provides valuable insights and tools for the continued advancement of AI. It underscores the importance of interdisciplinary approaches in AI research and paves the way for more intelligent, efficient, and user-aligned AI systems in the future.

5. Conclusions

Bibliography

- [1] Mohamed Afham, Salman Khan, Muhammad Haris Khan, Muzammal Naseer, and Fahad Shahbaz Khan. Rich semantics improve few-shot learning. *arXiv preprint arXiv:2104.12709*, 2021. [2.1](#)
- [2] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019. [4.2](#)
- [3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. [3.1](#), [3.2](#), [4.2](#), [4.4](#)
- [4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. [2.1](#)
- [5] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 33:9758–9770, 2020. [2.1](#)
- [6] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. [3.1](#), [4.1](#)
- [7] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*,

2023. [3.1](#)
- [8] Peyman Bateni, Raghav Goyal, Vaden Masrani, Frank Wood, and Leonid Sigal. Improved few-shot visual classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14493–14502, 2020. [2.1](#)
- [9] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3: 1137–1155, 2003. [4.2](#), [4.3](#)
- [10] Lorenzo Bertolini, Julie Weeds, and David Weir. Testing large language models on compositionality and inference with phrase-level adjective-noun entailment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4084–4100, 2022. [4.2](#)
- [11] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Joyce Zhuang, Juntang and Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiang, and Aditya Ramesh. Improving image generation with better captions. *Note on Dalle-3*, 2023. ([document](#)), [3.1](#), [3.3](#), [3.6](#), [3.3](#)
- [12] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. [2.1](#)
- [13] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101—mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014. [??](#)
- [14] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. [3.5](#), [??](#)
- [15] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019. [4.2](#)
- [16] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors,

- Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf>. 2, 2.1
- [17] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3.2, 4.1
- [18] Gemma Calvert, Edward Bullmore, M.J. Brammer, Ruth Campbell, Steven Williams, Philip McGuire, Peter Woodruff, S.D. Iversen, and Anthony David. Activation of auditory cortex during silent lipreading. *science*, 276(5312), 593–596. *Science (New York, N.Y.)*, 276:593–6, 05 1997. 2.1
- [19] Cătălina Cangea, Petar Veličković, and Pietro Lio. Xflow: Cross-modal deep neural networks for audiovisual classification. *IEEE Transactions on Neural Networks and Learning Systems*, 31(9):3711–3720, 2019. 2.1
- [20] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 2.1
- [21] Paola Cascante-Bonilla, Khaled Shehada, James Seale Smith, Sivan Doveh, Donghyun Kim, Rameswar Panda, Gül Varol, Aude Oliva, Vicente Ordonez, Rogerio Feris, et al. Going beyond nouns with vision & language models using synthetic data. *arXiv preprint arXiv:2303.17590*, 2023. 4.2
- [22] Yekun Chai, Shuohuan Wang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Clip-tuning: Towards derivative-free prompt learning with a mixture of rewards. *arXiv preprint arXiv:2210.12050*, 2022. 3.2
- [23] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 2.1
- [24] Jaemin Cho, Abhay Zala, and Mohit Bansal. Visual programming for text-to-image generation and evaluation. *arXiv preprint arXiv:2305.15328*, 2023. 3.7
- [25] Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou,

- Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416, 2022. 4.2, 1, 4.5
- [26] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. ??
- [27] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. ??
- [28] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. Why can gpt learn in-context? language models secretly perform gradient descent as meta optimizers. *arXiv preprint arXiv:2212.10559*, 2022. 3.2, 3.3
- [29] Béatrice Daille. *Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques*. PhD thesis, Ph. D. thesis, Université Paris 7, 1994. 4.6
- [30] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. (document), 2, 2.1, 2.5, 2.4, ??, 3.1, 3.4, 3.5, ??, 4.1, 4.2, 4.5
- [31] Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric P Xing, and Zhiting Hu. Rlprompt: Optimizing discrete text prompts with reinforcement learning. *arXiv preprint arXiv:2205.12548*, 2022. 2.1, 3.2
- [32] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4.2
- [33] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 2.1
- [34] Guneet S Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. *arXiv preprint arXiv:1909.02729*, 2019. 2, 2.1
- [35] Shizhe Diao, Xuechun Li, Yong Lin, Zhichao Huang, and Tong Zhang. Black-box prompt learning for pre-trained language models. *arXiv preprint arXiv:2201.08531*, 2022. 3.2
- [36] Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. Why is winoground hard? investigating failures in visuolinguistic composition-

- ality. *arXiv preprint arXiv:2211.00768*, 2022. 4.5
- [37] Sivan Doveh, Assaf Arbelle, Sivan Harary, Rameswar Panda, Roei Herzig, Eli Schwartz, Donghyun Kim, Raja Giryes, Rogerio Feris, Shimon Ullman, et al. Teaching structured vision&language concepts to vision&language models. *arXiv preprint arXiv:2211.11733*, 2022. 4.2
- [38] Sivan Doveh, Assaf Arbelle, Sivan Harary, Amit Alfassy, Roei Herzig, Donghyun Kim, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, et al. Dense and aligned captions (dac) promote compositional reasoning in vl models. *arXiv preprint arXiv:2305.19595*, 2023. 4.2
- [39] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 3.1
- [40] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022. 4.5
- [41] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. ??
- [42] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 2, 2.1
- [43] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*, 2023. 4.2, 1
- [44] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3-4):163–352, 2022. 3.2
- [45] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. (document), 2, 2.2, 2.1, 2.4, 2.2, 2.3
- [46] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020. 2.1
- [47] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade

- Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017. doi: 10.1109/ICASSP.2017.7952261. [2.4](#)
- [48] Eleanor J Gibson. Principles of perceptual learning and development. 1969. ([document](#)), [2.1](#), [2](#)
- [49] Rohit Girdhar and Deva Ramanan. Attentional pooling for action recognition. *Advances in neural information processing systems*, 30, 2017. [2.1](#)
- [50] Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, et al. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*, 2021. [2.1](#)
- [51] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. [3.1](#), [4.1](#), [4.2](#)
- [52] Chengcheng Guo, Bo Zhao, and Yanbing Bai. Deepcore: A comprehensive library for coreset selection in deep learning. In *Database and Expert Systems Applications: 33rd International Conference, DEXA 2022, Vienna, Austria, August 22–24, 2022, Proceedings, Part I*, pages 181–195. Springer, 2022. [4.9](#)
- [53] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. *arXiv preprint arXiv:2211.11559*, 2022. [3.2](#)
- [54] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio, 2021. ([document](#)), [2.1](#), [2](#), [2.1](#), [2.4](#), [2.4](#)
- [55] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Esresne(x)t-fbsp: Learning robust time-frequency transformation of audio, 2021. [2.4](#)
- [56] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *Proceedings of the IEEE international conference on computer vision*, pages 3018–3027, 2017. [2](#), [2.1](#)
- [57] Adi Haviv, Jonathan Berant, and Amir Globerson. Bertese: Learning to speak to bert. *arXiv preprint arXiv:2103.05327*, 2021. [2.1](#)
- [58] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [2.3](#)
- [59] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–

- 9738, 2020. [2](#), [2.1](#)
- [60] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. [2](#), [2.1](#)
- [61] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification, 2017. [3.4](#), [??](#)
- [62] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. [??](#)
- [63] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021. [2.5](#)
- [64] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. [2.5](#)
- [65] Christian Andreas Henning and Ralph Ewerth. Estimating the information gap between textual and visual representations. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 14–22, 2017. [4.6](#)
- [66] Roei Herzig, Alon Mendelson, Leonid Karlinsky, Assaf Arbelle, Rogerio Feris, Trevor Darrell, and Amir Globerson. Incorporating structured representations into pretrained vision & language models using scene graphs. *arXiv preprint arXiv:2305.06343*, 2023. [4.2](#)
- [67] Jack Hessel and Alexandra Schofield. How effective is bert without word ordering? implications for language understanding and data privacy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 204–211, 2021. [4.2](#)
- [68] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. [3.7](#)
- [69] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. [3.3](#)
- [70] Danfeng Hong, Naoto Yokoya, Gui-Song Xia, Jocelyn Chanussot, and Xiao Xi-

- ang Zhu. X-modalnet: A semi-supervised deep cross-modal network for classification of remote sensing data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 167:12–23, 2020. 2.1
- [71] Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420, 2017. 3.4
- [72] Or Honovich, Uri Shaham, Samuel R Bowman, and Omer Levy. Instruction induction: From few examples to natural language task descriptions. *arXiv preprint arXiv:2205.10782*, 2022. 3.2
- [73] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. (document), 2, 2.1, 2.2, 3.1
- [74] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *arXiv preprint arXiv:2306.14610*, 2023. 4.1, 4.2, 4.3, 4.4
- [75] Edward Hu, Yelong Shen, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 3.1
- [76] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*, 2023. 3.7
- [77] Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022. 2.1
- [78] Yufeng Huang, Jiji Tang, Zhuo Chen, Rongsheng Zhang, Xinfeng Zhang, Weijie Chen, Zeng Zhao, Tangjie Lv, Zhipeng Hu, and Wen Zhang. Structure-clip: Enhance multi-modal language representations with structure knowledge. *arXiv preprint arXiv:2305.06152*, 2023. 4.2
- [79] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL <https://doi.org/10.5281/zenodo.5143773>. 3.1
- [80] Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. Patching open-vocabulary models by interpolating weights. *arXiv preprint arXiv:2208.05592*, 2022. 2

- [81] Ray Jackendoff. On beyond zebra: The relation of linguistic and visual information. *Cognition*, 26(2):89–114, 1987. [2](#)
- [82] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. [4.2](#)
- [83] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. [2.1](#), [2.2](#)
- [84] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*, 2022. [2.1](#)
- [85] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020. [2.1](#)
- [86] Thorsten Joachims et al. Transductive inference for text classification using support vector machines. In *Icml*, volume 99, pages 200–209, 1999. [2](#), [2.1](#)
- [87] Armand Joulin, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision*, pages 67–84. Springer, 2016. [2.1](#)
- [88] Amita Kamath, Jack Hessel, and Kai-Wei Chang. Text encoders are performance bottlenecks in contrastive vision-language models. *arXiv preprint arXiv:2305.14897*, 2023. [4.1](#)
- [89] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019. [4.8](#)
- [90] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*, 2019. [4.2](#), [4.8](#), [4.9](#)
- [91] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. [2.1](#)
- [92] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural*

- information processing systems*, 35:22199–22213, 2022. 3.1
- [93] Stephen M. Kosslyn, Giorgio Ganis, and William L. Thompson. 3Multimodal images in the brain. In *The neurophysiological foundations of mental and motor imagery*. Oxford University Press, 01 2010. ISBN 9780199546251. doi: 10.1093/acprof:oso/9780199546251.003.0001. URL <https://doi.org/10.1093/acprof:oso/9780199546251.003.0001>. 2.1
- [94] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. ??
- [95] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. ??
- [96] Patricia K Kuhl and Andrew N Meltzoff. The intermodal representation of speech in infants. *Infant behavior and development*, 7(3):361–381, 1984. (document), 2.1, 2
- [97] Jet-Tsyn Lee, Danushka Bollegala, and Shan Luo. “touching to see” and “seeing to feel”: Robotic cross-modal sensory data generation for visual-tactile perception. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4276–4282. IEEE, 2019. 2.1
- [98] Li, Andreeto, Ranzato, and Perona. Caltech 101, Apr 2022. 3.5, ??
- [99] Ang Li, Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. Learning visual n-grams from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4183–4192, 2017. 2.1
- [100] Jiwei Li and Dan Jurafsky. Mutual information and diverse decoding improve neural machine translation, 2016. 4.6, 4.6, 4.8
- [101] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1014. URL <https://aclanthology.org/N16-1014>. 4.6, 4.6, 4.8
- [102] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 4.2
- [103] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping

- language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. ([document](#)), [4.1](#), [4.1](#), [4.2](#), [4.4](#), [4.4](#), [4.2](#), [4.4](#), [4.5](#), [4.8](#)
- [104] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. [2.1](#)
- [105] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. [3.1](#), [3.2](#), [4.1](#), [4.1](#), [4.2](#), [4.4](#), [4.5](#)
- [106] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. [2.1](#)
- [107] Shuang Li, Yilun Du, Joshua B Tenenbaum, Antonio Torralba, and Igor Mordatch. Composing ensembles of pre-trained models via iterative consensus. *arXiv preprint arXiv:2210.11522*, 2022. [3.2](#)
- [108] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020. [3.1](#)
- [109] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020. [2.1](#)
- [110] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. [3.2](#)
- [111] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. [2.1](#)
- [112] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. [3.1](#), [4.1](#), [4.2](#), [4.5](#), [4.7](#)
- [113] Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Revisiting the role of language priors in vision-language models. *arXiv preprint arXiv:2306.01879*, 2023. [3.7](#)

- [114] Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramana. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. *arXiv preprint arXiv:2301.06267*, 2023. [3.1](#), [3.2](#), [3.4](#), [??](#), [4.2](#), [4.5](#)
- [115] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021. [2.1](#), [2.5](#)
- [116] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023. [3.2](#)
- [117] Vivian Liu. Beyond text-to-image: Multimodal prompts to explore generative ai. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2023. [3.6](#)
- [118] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *arXiv:2103.10385*, 2021. [2](#)
- [119] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. [4.2](#)
- [120] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *arXiv preprint arXiv:2304.09842*, 2023. [3.2](#)
- [121] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5206–5215, 2022. [2](#), [2.1](#), [2.3](#), [2.5](#)
- [122] Shan Luo, Wenzhen Yuan, Edward Adelson, Anthony G Cohn, and Raul Fuentes. Vitac: Feature sharing between vision and tactile sensing for cloth texture recognition. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2722–2727. IEEE, 2018. [2.1](#)
- [123] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? *arXiv preprint arXiv:2212.07796*, 2022. [4.1](#), [4.2](#), [4.4](#)
- [124] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. [3.1](#)

- [125] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. [3.4](#), [??](#)
- [126] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. [??](#)
- [127] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021. [4.9](#)
- [128] Andrew N Meltzoff and Richard W Borton. Intermodal matching by human neonates. *Nature*, 282(5737):403–404, 1979. ([document](#)), [2.1](#), [2](#)
- [129] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022. [3.1](#), [3.2](#), [3.4](#), [??](#)
- [130] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2021. [4.9](#)
- [131] Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. Reframing instructional prompts to gptk’s language. *arXiv preprint arXiv:2109.07830*, 2021. [3.2](#)
- [132] Melanie Mitchell, John Holland, and Stephanie Forrest. When will a genetic algorithm outperform hill climbing. *Advances in neural information processing systems*, 6, 1993. [3.2](#)
- [133] Jesse Mu, Percy Liang, and Noah Goodman. Shaping visual representations with language for few-shot classification. *arXiv preprint arXiv:1911.02683*, 2019. [2](#), [2.1](#), [2.3](#), [2.3](#)
- [134] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, pages 529–544. Springer, 2022. [2.1](#)
- [135] Bence Nanay. Multimodal mental imagery. *Cortex*, 105:125–136, 2018. ([document](#)), [2.1](#), [2](#), [2.1](#)
- [136] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008. [??](#)
- [137] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. [??](#)
- [138] Changdae Oh, Hyeji Hwang, Hee-young Lee, YongTaek Lim, Geunyoung Jung,

- Jiyoung Jung, Hosik Choi, and Kyungwoo Song. Blackvip: Black-box visual prompting for robust transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24224–24235, 2023. [3.2](#), [3.3](#), [3.7](#)
- [139] OpenAI. Gpt-4 technical report. 2023. ([document](#)), [3.1](#), [3.2](#), [3.3](#), [3.6](#), [3.1](#), [3.3](#), [4.2](#), [4.9](#)
- [140] Yassine Ouali, Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. Black box few-shot adaptation for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15534–15546, 2023. [3.2](#), [3.3](#)
- [141] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022. ([document](#)), [3.1](#), [3.2](#), [3.3](#), [3.1](#)
- [142] Frederik Pahde, Main Nabi, Tassila Klein, and Patrick Jahnichen. Discriminative hallucination for multi-modal few-shot learning. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 156–160. IEEE, 2018. [2.1](#)
- [143] Frederik Pahde, Mihai Puscas, Tassilo Klein, and Moin Nabi. Multimodal prototypical networks for few-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2644–2653, 2021. [2.1](#)
- [144] Isabel Papadimitriou, Richard Futrell, and Kyle Mahowald. When classifying grammatical role, bert doesn’t care about word order... except when it matters. *arXiv preprint arXiv:2203.06204*, 2022. [4.2](#)
- [145] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. [??](#)
- [146] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. [??](#)
- [147] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015. [2](#), [2.4](#)
- [148] Archiki Prasad, Peter Hase, Xiang Zhou, and Mohit Bansal. Grips: Gradient-free, edit-based instruction search for prompting large language models. *arXiv preprint arXiv:2203.07281*, 2022. [2.1](#), [3.2](#)

- [149] Sarah Pratt, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. *arXiv preprint arXiv:2209.03320*, 2022. [3.1](#), [3.2](#)
- [150] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5822–5830, 2018. [2](#), [2.1](#)
- [151] R Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107, 2005. [2](#)
- [152] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. 2019. [4.2](#)
- [153] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. ([document](#)), [3.1](#), [3.2](#), [3.3](#), [3.1](#), [??](#), [3.1](#), [4.1](#), [4.2](#)
- [154] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 2021. ([document](#)), [2.1](#), [2](#), [2.1](#), [2.2](#), [2.3](#), [2.4](#), [2.5](#), [??](#)
- [155] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. [3.1](#)
- [156] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016. [2](#), [2.1](#)
- [157] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. [2.5](#)
- [158] François Role and Mohamed Nadif. Handling the impact of low frequency events on co-occurrence based measures of word similarity. In *Proceedings of the international conference on Knowledge Discovery and Information Retrieval (KDIR-2011)*. Scitepress, pages 218–223, 2011. [4.6](#)
- [159] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. [3.1](#), [3.6](#)
- [160] Stuart J Russell. *Artificial intelligence a modern approach*. Pearson Education,

- Inc., 2010. 3.2, 3.4
- [161] Timo Schick and Hinrich Schütze. Exploiting cloze questions for few-shot text classification and natural language inference. *Computing Research Repository*, arXiv:2001.07676, 2020. URL <http://arxiv.org/abs/2001.07676>. 2.1
- [162] Timo Schick and Hinrich Schütze. It’s not just size that matters: Small language models are also few-shot learners. *Computing Research Repository*, arXiv:2009.07118, 2020. URL <http://arxiv.org/abs/2009.07118>. 2.1
- [163] Lauren A Schmidt. *Meaning and compositionality as statistical induction of categories and constraints*. PhD thesis, Massachusetts Institute of Technology, 2009. 2
- [164] Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001. 2.2
- [165] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 3.1, 3.4, 4.2, 4.4
- [166] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 3.1, 3.2, 4.2
- [167] Eli Schwartz, Leonid Karlinsky, Rogerio Feris, Raja Giryes, and Alex Bronstein. Baby steps towards few-shot learning with multiple semantics. *Pattern Recognition Letters*, 160:142–147, 2022. 2.1
- [168] Alexander Shapiro. Monte carlo sampling methods. *Handbooks in operations research and management science*, 10:353–425, 2003. 4.4
- [169] Sheng Shen, Chunyuan Li, Xiaowei Hu, Yujia Xie, Jianwei Yang, Pengchuan Zhang, Zhe Gan, Lijuan Wang, Lu Yuan, Ce Liu, et al. K-lite: Learning transferable visual models with external knowledge. *Advances in Neural Information Processing Systems*, 35:15558–15573, 2022. 3.2
- [170] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023. 3.2
- [171] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020. 2.1,

3.2

- [172] Aman Shrivastava, Ramprasaath R Selvaraju, Nikhil Naik, and Vicente Ordonez. Clip-lite: information efficient visual representation learning from textual annotations. *arXiv preprint arXiv:2112.07133*, 2021. 4.6
- [173] Harman Singh, Pengchuan Zhang, Qifan Wang, Mengjiao Wang, Wenhan Xiong, Jingfei Du, and Yu Chen. Coarse-to-fine contrastive learning in image-text-graph space for improved vision-language compositionality. *arXiv preprint arXiv:2305.13812*, 2023. 4.2
- [174] Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *arXiv preprint arXiv:2104.06644*, 2021. 4.2
- [175] Linda Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artificial life*, 11(1-2):13–29, 2005. 2
- [176] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 2, 2.1, 2.2
- [177] Haoyu Song, Li Dong, Wei-Nan Zhang, Ting Liu, and Furu Wei. Clip models are few-shot learners: Empirical studies on vqa and visual entailment. *arXiv preprint arXiv:2203.07190*, 2022. 2.1
- [178] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. ??, 3.5, ??
- [179] Tianxiang Sun, Zhengfu He, Hong Qian, Yunhua Zhou, Xuan-Jing Huang, and Xipeng Qiu. Bbtv2: Towards a gradient-free future with large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3916–3930, 2022. 3.2
- [180] Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. Black-box tuning for language-model-as-a-service. In *International Conference on Machine Learning*, pages 20841–20855. PMLR, 2022. 3.2
- [181] Ajinkya Tejanekar, Maziar Sanjabi, Bichen Wu, Saining Xie, Madian Khabza, Hamed Pirsiavash, and Hamed Firooz. A fistful of words: Learning transferable visual models from bag-of-words supervision. *arXiv preprint arXiv:2112.13884*, 2021. 4.2
- [182] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF*

- Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. (document), 3.1, 3.6, 3.3, 4.1, 4.1, 4.2, 4.5
- [183] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 2.1
- [184] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL <https://proceedings.neurips.cc/paper/2016/file/90e1357833654983612fb05e3ec9148c-Paper.pdf>. 2.1
- [185] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019. 2.5
- [186] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 3.1
- [187] Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Equivariant similarity for vision-language foundation models. *arXiv preprint arXiv:2303.14465*, 2023. 4.1, 4.2, 4.5
- [188] Xudong Wang, Zhirong Wu, Long Lian, and Stella X Yu. Debaised learning from naturally imbalanced pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14647–14657, 2022. 2.1
- [189] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Growing a brain: Fine-tuning by increasing model capacity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2471–2480, 2017. 2.1
- [190] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7278–7286, 2018. 2
- [191] Zeyu Wang, Berthy Feng, Karthik Narasimhan, and Olga Russakovsky. Towards unique and informative captioning of images. In *European Conference on Computer Vision (ECCV)*, 2020. 4.6, 4.6
- [192] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022. 3.6

- [193] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 3.1
- [194] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. *arXiv preprint arXiv:2109.01903*, 2021. <https://arxiv.org/abs/2109.01903>. (document), 3.1, 3.4, ??, 3.1
- [195] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022. (document), 2, 2.1, 2.4, 2.2, 2.2, 2.3, ??, 2.3, ??, 2.2, 2.5, 4.2
- [196] Wenhao Wu, Zhun Sun, and Wanli Ouyang. Transferring textual knowledge for visual recognition. *arXiv preprint arXiv:2207.01297*, 2022. 2.1
- [197] Xindi Wu, Zhiwei Deng, and Olga Russakovsky. Multimodal dataset distillation for image-text retrieval. *arXiv preprint arXiv:2308.07545*, 2023. 4.8, 4.9
- [198] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018. 2.3
- [199] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. ??
- [200] Jianxiong Xiao, Krista A. Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *Int. J. Comput. Vision*, 119(1):3–22, aug 2016. ISSN 0920-5691. doi: 10.1007/s11263-014-0748-y. URL <https://doi.org/10.1007/s11263-014-0748-y>. ??
- [201] Chen Xing, Negar Rostamzadeh, Boris Oreshkin, and Pedro O O Pinheiro. Adaptive cross-modal few-shot learning. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 2.1, 2.3, 2.3
- [202] Yinghui Xing, Qirui Wu, De Cheng, Shizhou Zhang, Guoqiang Liang, and Yanning Zhang. Class-aware visual prompt tuning for vision-language pre-trained model. *arXiv preprint arXiv:2208.08340*, 2022. 2.1
- [203] Hanwei Xu, Yujun Chen, Yulun Du, Nan Shao, Yanggang Wang, Haiyu Li, and

- Zhilin Yang. Gps: Genetic prompt search for efficient few-shot learning. *arXiv preprint arXiv:2210.17041*, 2022. 3.2
- [204] Ting Yao, Tao Mei, and Chong-Wah Ngo. Co-reranking by mutual reinforcement for image search. In *Proceedings of the ACM international conference on image and video retrieval*, pages 34–41, 2010. 4.6
- [205] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 3.1, 4.1, 4.1, 4.5, 4.7
- [206] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2.1, 3.1, 4.4
- [207] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc., 2021. URL <https://proceedings.neurips.cc/paper/2021/file/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Paper.pdf>. 4.2, 1
- [208] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bag-of-words models, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022. 4.1, 4.1, 4.2, 4.3, 4.4, 4.5
- [209] Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022. 3.2
- [210] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 2.1
- [211] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 833–842, 2021. 2.1
- [212] Haotian* Zhang, Pengchuan* Zhang, Xiaowei Hu, Yen-Chun Chen, Lianian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *arXiv preprint arXiv:2206.05836*, 2022. 2.1

- [213] Jeffrey O Zhang, Alexander Sax, Amir Zamir, Leonidas Guibas, and Jitendra Malik. Side-tuning: a baseline for network adaptation via additive side networks. In *European Conference on Computer Vision*, pages 698–714. Springer, 2020. [2.1](#)
- [214] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. ([document](#)), [2](#), [2.1](#), [2.4](#), [2.2](#), [2.3](#), [2.3](#), [??](#), [??](#), [2.1](#), [2.3](#), [??](#), [??](#), [2.5](#), [??](#), [??](#), [??](#), [??](#)
- [215] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. [4.2](#), [1](#), [4.5](#)
- [216] Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*, 2022. [4.1](#), [4.2](#), [4.4](#)
- [217] Tony Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, 2021. [4.4](#)
- [218] Mingkai Zheng, Xiu Su, Shan You, Fei Wang, Chen Qian, Chang Xu, and Samuel Albanie. Can gpt-4 perform neural architecture search? *arXiv preprint arXiv:2304.10970*, 2023. [3.2](#)
- [219] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. ([document](#)), [3.1](#), [3.2](#), [3.4](#), [3.5](#), [3.5](#), [3.1](#), [??](#), [3.1](#)
- [220] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. [2](#), [2.1](#), [2.5](#), [2.5](#)
- [221] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022. ([document](#)), [2](#), [2.1](#), [2.4](#), [2.3](#), [2.3](#), [??](#), [2.1](#), [2.3](#), [??](#), [2.2](#), [2.5](#), [2.5](#), [2.5](#), [2.7](#), [??](#), [??](#)
- [222] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*, 2022. [3.1](#), [3.2](#)
- [223] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. *arXiv preprint arXiv:2205.14865*, 2022. [2](#), [2.1](#), [??](#)